

# ABSTRACT

Title of Dissertation: DINOFLAGELLATE GENOMIC  
ORGANIZATION AND PHYLOGENETIC  
MARKER DISCOVERY UTILIZING DEEP  
SEQUENCING DATA

Gregory Scott Mendez,  
Doctor of Philosophy, 2016

Dissertation directed by: Professor Charles F. Delwiche,  
Cell Biology and Molecular Genetics

Dinoflagellates possess large genomes in which most genes are present in many copies. This has made studies of their genomic organization and phylogenetics challenging. Recent advances in sequencing technology have made deep sequencing of dinoflagellate transcriptomes feasible. This dissertation investigates the genomic organization of dinoflagellates to better understand the challenges of assembling dinoflagellate transcriptomic and genomic data from short read sequencing methods, and develops new techniques that utilize deep sequencing data to identify orthologous genes across a diverse set of taxa. To better understand the genomic organization of dinoflagellates, a genomic cosmid clone of the tandemly repeated gene Alcohol Dehydrogenase (AHD) was sequenced and analyzed. The organization of this clone was found to be counter to prevailing hypotheses of genomic organization in dinoflagellates. Further, a new non-canonical splicing motif was described that could greatly improve the automated modeling and annotation of genomic data. A custom

phylogenetic marker discovery pipeline, incorporating methods that leverage the statistical power of large data sets was written. A case study on Stramenopiles was undertaken to test the utility in resolving relationships between known groups as well as the phylogenetic affinity of seven unknown taxa. The pipeline generated a set of 373 genes useful as phylogenetic markers that successfully resolved relationships among the major groups of Stramenopiles, and placed all unknown taxa on the tree with strong bootstrap support. This pipeline was then used to discover 668 genes useful as phylogenetic markers in dinoflagellates. Phylogenetic analysis of 58 dinoflagellates, using this set of markers, produced a phylogeny with good support of all branches. The *Suessiales* were found to be sister to the *Peridinales*. The *Prorocentrales* formed a monophyletic group with the *Dinophysiales* that was sister to the *Gonyaulacales*. The *Gymnodinales* was found to be paraphyletic, forming three monophyletic groups. While this pipeline was used to find phylogenetic markers, it will likely also be useful for finding orthologs of interest for other purposes, for the discovery of horizontally transferred genes, and for the separation of sequences in metagenomic data sets.

DINOFLAGELLATE GENOMIC ORGANIZATION AND  
PHYLOGENETIC MARKER DISCOVERY UTILIZING  
DEEP SEQUENCING DATA

by

Gregory Scott Mendez

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2016

Advisory Committee:

Professor Professor Charles F. Delwiche, Chair  
Professor Tsvetan Bachvaroff  
Professor Priscila Chaverri  
Professor Najib El-Sayed  
Professor Charles Mitter  
Professor Eric Haag

© Copyright by  
Gregory Scott Mendez  
2016

# Table of Contents

<b>Table of Contents .....</b>	<b>ii</b>
<b>List of Figures.....</b>	<b>v</b>
<b>List of Tables .....</b>	<b>vi</b>
<b>List of Abbreviations .....</b>	<b>vii</b>
<b>1 Introduction.....</b>	<b>1</b>
1.1 Dinoflagellates .....	1
1.2 Dinoflagellate Genomes.....	2
1.3 Dinoflagellate Phylogeny.....	4
<b>2 Dinoflagellate Gene Structure and Intron Splice Sites in a Genomic Tandem Array .....</b>	<b>8</b>
2.1 Background.....	8
2.1.1 Nucleic Acid Isolation .....	12
2.2 Cosmid Library Construction and Screening.....	12
2.2.1 DNA Sequencing and Analysis .....	13
2.3 Results.....	14
2.3.1 Analysis of Cosmid Sequence .....	14
2.3.2 Map of Cosmid Sequence .....	14
2.3.3 Sequence Similarity Analysis .....	16
2.3.4 Intron Border Repeat (IRIB).....	19
2.4 Discussion.....	21
2.4.1 The Consensus Model.....	21
2.4.2 Conservation of Non-Coding Regions.....	23

2.4.3	Non-Canonical Splicing of Introns .....	25
<b>3</b>	<b>A Method of Screening Genomic and Transcriptomic Libraries for</b>	
	<b>Probable Orthologs: A Stramenopile Case Study.....</b>	<b>27</b>
3.1	Background.....	27
3.2	Implementation .....	28
3.2.1	Overview.....	28
3.2.2	Ortholog Search and Scoring.....	30
3.2.3	Constraint Tree Branch Length Outlier Analysis .....	33
3.2.4	Distance-Matrix Outlier Analysis.....	36
3.2.5	In-paralog and Sister Paralog Search.....	39
3.2.6	Preparing Inputs for Second Round of Sequence Searches .....	40
3.2.7	Tests of Second Round of Searches.....	41
3.2.8	Preparation of Input Sequences for Case Study.....	42
3.3	Results and Discussion .....	43
3.3.1	Case Study Input.....	43
3.3.2	Case Study Statistics.....	44
3.3.3	Case Study Phylogenetics.....	45
3.4	Conclusions.....	52
3.5	Availability of supporting data .....	53
<b>4</b>	<b>Application of a New Ortholog Detection Pipeline to the Discovery of</b>	
	<b>Phylogenetic Markers in Dinoflagellates .....</b>	<b>54</b>
4.1	Introduction.....	54
4.2	Materials and Methods.....	57

4.2.1	Taxon Selection and Transcriptomic Assembly .....	57
4.2.2	Query Sequence Preparation.....	58
4.2.3	Ortholog Identification.....	60
4.2.4	Phylogenetic Analysis.....	60
4.3	Results.....	60
4.3.1	Sequence Searches and Screening .....	60
4.3.2	Phylogenetic Analyses.....	61
4.4	Discussion.....	66
<b>5</b>	<b>Conclusions.....</b>	<b>69</b>
	<b>Appendix A – Supplemental Figures .....</b>	<b>73</b>
	<b>Appendix B – Supplemental Tables .....</b>	<b>77</b>
	<b>References.....</b>	<b>83</b>

## List of Figures

Figure 1 – Synthesis of Molecular and Morphological Studies.....	5
Figure 2 – Schematic of Alcohol Dehydrogenase Cosmid Insert.....	15
Figure 3 – Pairwise Identity Graphs for Intergenic Spacers.....	18
Figure 4 – Identical Repeated Intron Boundary.....	20
Figure 5 – Overview of Pipeline.....	29
Figure 6 – Flow Chart of Initial Ortholog Searches .....	32
Figure 7 – Flow Chart of Paralog Filtering Process .....	34
Figure 8 – Histogram and Cladogram of Long Branch Analysis .....	36
Figure 9 – Example of Distance Matrix Outlier Analysis Gene Tree.....	38
Figure 10 – Example In-Paralog and Sister-Paralog Analysis.....	40
Figure 11 – Stramenopile Nucleic Acid Maximum Likelihood Tree .....	48
Figure 12 – Stramenopile Amino Acid Maximum Likelihood Tree .....	50
Figure 13 – Dinoflagellate Nucleic Acid Maximum Likelihood Tree .....	64
Figure 14 – Dinoflagellate Amino Acid Maximum Likelihood Tree.....	65



## List of Tables

Table 1 – Species with uncertain identification.....	44
Table 2 – Assemblies used for Clustering .....	59

## List of Abbreviations

rDNA.....	ribosomal DNA
PCR .....	polymerase chain reaction
EST .....	expressed sequence tag
ADH .....	alcohol dehydrogenase
IRIB .....	Intron Border Repeat
BLAST .....	Basic Local Alignment Search Tool
CEGMA .....	Core Eukaryotic Gene Mapping Approach
BUSCO .....	Benchmarking Universal Single-Copy Orthologs
HMM .....	Hidden Markov Model
OTU .....	Operational Taxonomic Unit
HTML .....	Hyper Text Markup Language
MSAs .....	Multiple sequence alignments
HGT .....	horizontal gene transfer
MAD .....	Media Absolute Deviation
NCBI .....	The National Center for Biotechnology Information
MMETSP .....	The Marine Microbial Eukaryote Transcriptome Sequences Project
ORFs .....	open-reading frames
ML .....	Maximum-Likelihood

# 1 Introduction

## 1.1 Dinoflagellates

Dinoflagellates are bi-flagellate protists that possess such a wide variety of bizarre nuclear characteristics that they were once regarded as a "missing-link" between Prokaryotes and Eukaryotes. At the time, they were referred to as *Dinokaryota*- or *Mesokaryota*, and were interpreted as fourth domain of life unto themselves[1]. When molecular phylogenetic evidence for a variety of eukaryotes began to emerge in the late 1980s and early 1990s, it quickly became apparent that dinoflagellates were not only unambiguously of eukaryotic affinity, but that they were related to the parasitic apicomplexans and heterotrophic ciliates[2-4]. This group was named the *Alveolata*, named for the cortical alveoli underlying the cell membranes of the three members of the group, a diagnostic feature only noted after phylogenetic analyses revealed their close relationship[4]. While molecular phylogenetic methods proved critical in revealing the relationship of dinoflagellates to other eukaryotes, the unusual nuclear characteristics of the group continue to vex studies of the evolutionary relationships among dinoflagellates.

Dinoflagellates can be identified by the presence of two dimorphic flagella: a ribbon-like flagellum that beats to the cell's left and a rudder-like flagellum that beats posteriorly. The beating of these two dissimilar flagella give the cells a characteristic whirling motion as they swim that lends the group the stem of their name – *dinos* means "to whirl" in Greek. While most dinoflagellates are unicellular, filamentous (multicellular chains of cells) and coenocytic (multi-nucleate) forms are also known. Within the cortical alveoli of some dinoflagellates (thecate or armored species) are cellulosic plates conveying strength and shape to the cell. The arrangements of these plates were the basis for taxonomic studies of dinoflagellates until molecular methods

became widely available in the 1990s. Plate tabulation remains the gold standard for morphological taxonomy of most genera. However, the reliance on tabulation has resulted in a large and poorly understood clade of dinoflagellates, termed *Gymnodinales*, that lack this diagnostic feature (though a laborious procedure involving the stripping of the outer membrane can be used to investigate the arrangement of the cortical alveolae).

## 1.2 Dinoflagellate Genomes

Dinoflagellates possess many unusual characteristics pertaining to their genomes and nuclei. Three features in particular were the cause for the taxonomic confusion that placed dinoflagellates outside of the domain Eukaryota: chromosomes that remain permanently condensed throughout the cell cycle, the absence of protein associated with the DNA, and the fibrillar arched banded shape of the chromosomes, interpreted as a liquid crystal structure. As modern molecular techniques were employed, new genomic oddities have presented additional challenges to molecular phylogenetics. Four characteristics in particular make phylogenetic analysis in this group difficult:

1. ribosomal DNA (rDNA) is a poor phylogenetic marker, resulting in poorly supported clades that change greatly depending on the taxa selected for the analysis.
2. The mitochondrial and chloroplast genomes are small and unusual in structure
3. The nuclear genome size averages 20X the human genome.
4. Most, if not all, genes occur in large complex gene families

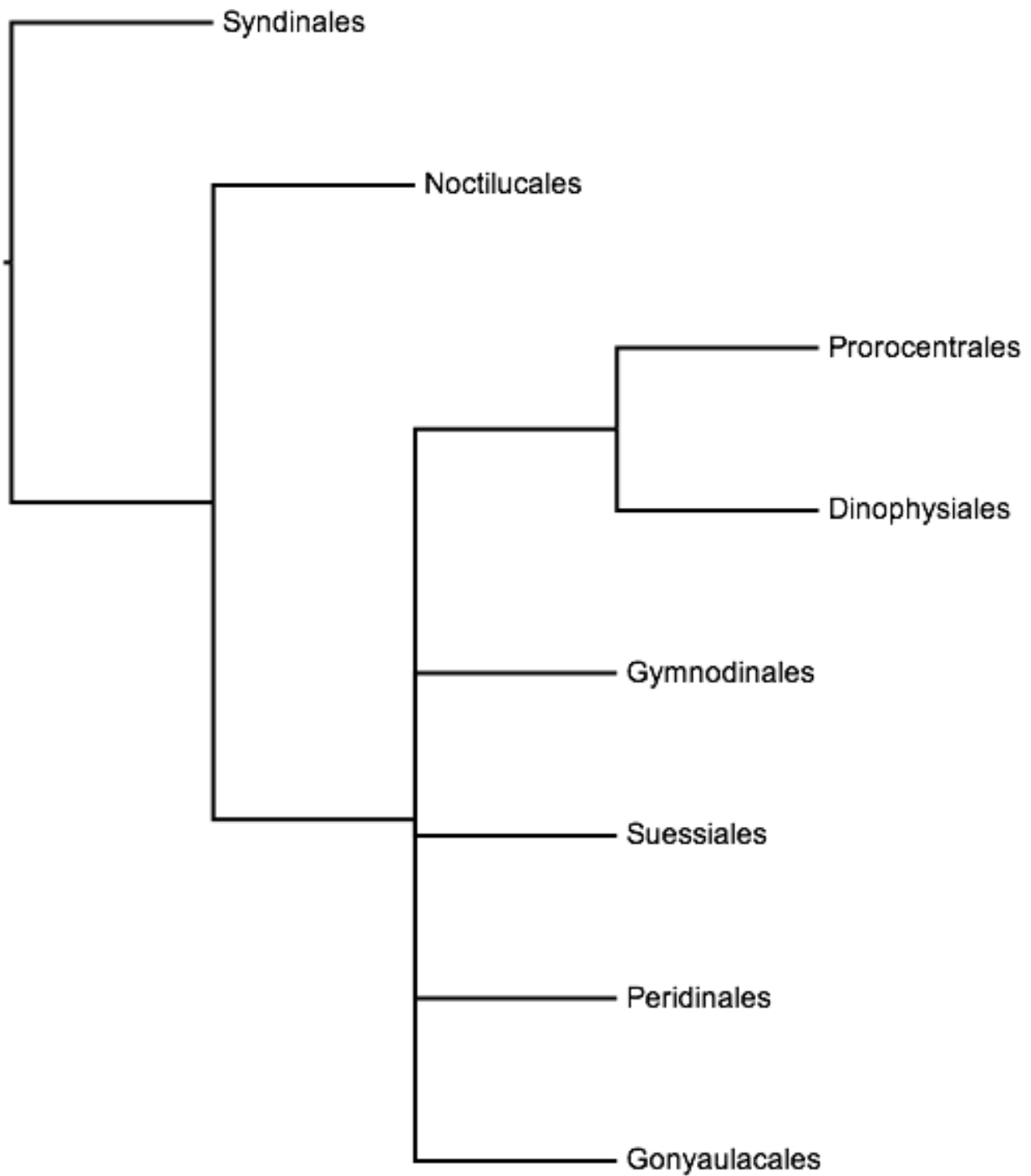
As with many organisms, rDNA trees were among the first to become available for dinoflagellates [3], and remain the basis for most modern analyses[5-9]. These analyses, however, have many areas of incongruence, failing to resolve groups strongly supported by morphological analyses, and changing depending on the taxa that are included. While

complementing rDNA data with data from mitochondrial or chloroplast genes would be a natural next step, both plastid and mitochondrial genomes have undergone a massive transfer of genes to the nuclear genome[10], making ortholog identification difficult. The plastid genome has been reduced to a handful of genes on 1-2 gene mini-circles that replicate using rolling circle intermediates[11,12]. Among the dinoflagellate class *Dinophyceae*, a monophyletic group of dinoflagellates that possess a dinokaryon, genome sizes vary by as much as two orders of magnitude. The average size of a dinophycean genome is more than 10x larger than the human genome, and some species can be as large as 100x the size of the human genome[13].

*Prorocentrum micans*, for example, measures 225pg of DNA per haploid cell, a genome over 70x the size of the human genome[13]. Using modern sequencing technology, it would cost well over \$100 million dollars to sequence the genome of *Prorocentrum micans* (ignoring assembly and analysis costs). Although early physiochemical analyses of the dinoflagellate genome had hinted at an unusual structure and relatively little repetition[14], it wasn't until DNA sequencing methods were employed in the 1990s that it was noted that dinoflagellate genes are organized as tandem repeats [15-18]. While little genomic data exists, it appears that most dinophycean genes are members of large heterogeneous families [19]. The heterogeneity of these gene families makes targeted polymerase chain reaction (PCR) sequencing methods and identifying orthologs challenging. Recent genomic surveys of dinoflagellates of the genus *Symbiodinium*, an endosymbiotic genus with remarkably small genomes for dinophycean dinoflagellates, have revealed little regarding the large heterogeneous gene families so characteristic of dinophycean dinoflagellates[20-22]. Collectively the unusual characteristics of the dinoflagellate genome interfere with most current methods of molecular phylogenetic analysis.

### 1.3 Dinoflagellate Phylogeny

While further study over the past two decades has solidified the position of dinoflagellates within Alveolata, resolution of the relationships among the dinoflagellates has seen little progress. Taylor's synthesis of phylogenetic and morphological data, identifies six major groups: *Gymnodinales*, *Peridinales*, *Suessiales*, *Gonyaulacales*, *Prorocentrales*, and *Dinophysiales*[23]. These initial six clades of dinoflagellates described by Taylor[23] all unambiguously exhibit a dinokaryon and make up the dinophyceae. Later analyses by Taylor [24] and Fensome[25], added the *Syndinales* and *Noctilucales*, which were thought to lack a dinokaryon all the time or only part of the time (respectively). Study of the *Syndinales* species *Amoebophrya* would later revise this understanding when a dinokaryon was observed during specific times during its life cycle[26]. While the *Syndinales* and *Noctilucales* are considered members of *dinokaryota*, their cryptic expression of the dinokaryon has placed them outside of the dinophyceae in syntheses of phylogenetic and morphological data[24,27](Figure 1). Of these eight major groups only two, the *Suessiales* and *Dinophysiales*, reproducibly form monophyletic groups in molecular phylogenetic analyses.



**Figure 1 – Synthesis of Molecular and Morphological Studies**

A synthesis of data from molecular and morphological studies. Note the eight major groups identified: Syndinales, Gymnodinales, Noctilucales, Peridinales, Suessiales, Gonyaulacales, Prorocentrales, and Dinophysioales. [23,24,27]

Phylogenetic hypotheses of dinoflagellate in-group relationships differ considerably depending on the taxa selected for the study and the genes used to find the trees. Branch lengths for the majority of the phylogenies are typically very short (including short internal branches), but a smattering of species with long branches are present. These long branch species are frequently located in different positions on the tree depending on the analysis performed. For example, *Cryptothecodinium cohnii*, widely regarded as a basal dinophycean, is found in the Gonyaulacales in one analysis and among the earliest branching dinophyceans in another analysis, seemingly based upon the selection of outgroups.

Difficulties in culturing dinoflagellates and in amplifying genes of interest from dinoflagellates has meant that phylogenetic studies utilizing more than one gene remain rare, and taxon sampling continues to hamper analyses. While taxon sampling is improving, most analyses still utilize only rDNA genes (usually the small and large subunit of the nuclear operon; SSU or LSU). Due to the impracticality of sequencing a whole dinoflagellate nuclear genome, researchers next turned their attention to the creation of expressed sequence tag (EST) libraries. EST libraries are produced using Sanger sequenced individual reads from a cDNA library. These libraries have underpinned many of the unusual findings previously discussed, but their utility in comparative biological studies such as phylogenetics has been hamstrung by the low diversity of species represented among the larger libraries. Of the dinoflagellate EST libraries over 10,000 sequences, only two of the eight major dinoflagellate lineages recognized by Taylor [24] are represented (Figure 1). Even when libraries with as few as 500 reads are considered, only four of the eight lineages are represented. A recent landmark study of dinoflagellate phylogeny sought to employ more recent Illumina RNA sequencing together with data from existing EST libraries[28]. While Bachvaroff et al utilized all publicly available data and nearly doubled that



data with newly sequenced taxa, it was still limited by taxon sampling, most notably lacking any taxa from the *Peridinales*, *Noctilucales*, or *Dinophysiales* [28].

The work presented here first investigates some of the hypotheses surrounding the unusual genomic organization of dinophycean dinoflagellates. Specifically, questions relevant to ortholog identification and the assembly of dinoflagellate sequences from modern short-read high-throughput sequencing methods are investigated in the context of one tandem gene array (Chapter 2). Next, deep sequencing data available as a result of advances in high-throughput sequencing is leveraged in new ways to overcome some of the limitations of existing ortholog identification tools that are central to modern phylogenetic pipelines. A custom ortholog identification and paralog screening pipeline was developed and tested (Chapter 3). This pipeline builds upon existing ortholog identification techniques, but uniquely takes advantage of statistics of the aggregated dataset to inform the selection of orthologs. Lastly, new publicly available dinoflagellate transcriptome data were analyzed using this pipeline to generate a set of 668 genes from 58 taxa representing the most comprehensive phylogenetic hypothesis of dinoflagellates to date, including all eight of the major clades of dinoflagellates (Chapter 4). Collectively this work provides insights not only to the evolution of this ecologically and economically important group of organisms, but also provides bioinformatics tools that may improve ortholog identification, automated genome annotation, and metagenomics studies.

## 2 Dinoflagellate Gene Structure and Intron Splice Sites in a Genomic Tandem Array

### 2.1 Background

Dinoflagellates are biflagellate protists that can be found in most of the world's aquatic environments. Depending upon the species, they play the diverse environmental roles of predators, prey, parasites, symbionts, and primary producers, with many showing plasticity of nutritive mode.

Dinoflagellates possess such a wide variety of unique nuclear characteristics that they were, until the 1990s, widely regarded as a "missing-link" between Prokaryotes and Eukaryotes. They were referred to as Dino- or *Mesokaryota*, and sometimes viewed as a fourth domain of life unto themselves [1]. These perplexing nuclear characteristics include large genomes, modified DNA bases, permanently condensed liquid-crystalline cholesteric-like chromosomes, a lack of nucleosomes, highly duplicated genes found in tandem arrays, a gene organization lacking typical eukaryotic conserved motifs, and a massive transfer of plastid genes to the nuclear genome [10,13,17,19,29-32]. The application of phylogenetic methods and molecular systematic data revealed that dinoflagellates reside firmly in the crown of the eukaryotes, among the Alveolates rather than belonging to a unique domain of life or even a basal lineage of eukaryotes [4,33].

Perhaps the most striking feature of a dinoflagellate cell is the large nucleus containing permanently condensed chromosomes. Dinoflagellate genome sizes vary by as much as two orders of magnitude, but the smallest dinoflagellate genome yet measured belongs to the endosymbiotic *Symbioninium* spp. with 1.5 pg per haploid cell, approximately half the size of the

human genome [13]. The average size of a Dinoflagellate genome is more than 10x larger than the human genome, and some species can be as large as 100x the size of the human genome [13]. These genomes are prohibitively large to analyze with limited resources using current sequencing and assembly technology. At present, the most complete genomic data published are from the diminutive *Symbiodinium minutum* genome, which represents approximately 41% of the genome in 33,815 contigs across 21,898 scaffolds. Despite its incompleteness and fragmentation, the genome survey represents the best look at a dinoflagellate genome to date. The intractability of completing an assembled dinoflagellate genome has meant that most dinoflagellate sequences have been generated primarily using two methods: shotgun sequencing of transcriptome libraries (e.g., EST sequencing), and PCR. Sequencing of mRNA is valuable, but by definition carries essentially no information about genome structure, and PCR-based methods depend upon flanking conserved primers, which imposes constraints on the insights that can be obtained from them.

*Cryptothecodinium cohnii* is a heterotrophic marine dinoflagellate with uncertain phylogenetic affinity; in some analyses it is placed in the crown of the Gonyaulacoid lineage [34-37], while other analyses find it placed with more basal dinoflagellate lineages [38-40]. *C. cohnii* has been used in the industrial manufacture of omega-3 fatty acids for fortification of infant formula [41]. Among dinoflagellates, *C. cohnii* is a relatively facile organism, capable of culture in either liquid or solid media, axenic culture, and growth in a specialized growth medium that promotes accelerated growth such that cultures reach late log phase 4-10X faster than other dinoflagellates. The genome size of *C. cohnii* is a third the size of the average dinoflagellate genome at 3.8 pg per haploid cell [42]. Many important discoveries have been made using *C. cohnii*, including the discovery of rare bases in dinoflagellate DNA, mutagenesis and breeding studies, and the low

protein content of dinoflagellate chromosomes (Rae 1973; Tuttle and Loeblich III 1974; Rizzo and Noodén 1972). The gene alcohol dehydrogenase (ADH) was targeted in the present study because it was found to be highly expressed in a *C. cohnii* cDNA library [43]

Prior to a genome survey of *Symbiodinium minutum*, understanding of dinoflagellate genomic organization was almost exclusively based on a small number of publications comprising just 11 sequences of 6 genes from 8 species [15-18,32,44-50]. Despite their paucity, these data led to an understanding of dinoflagellate genomic organization that is different from that of other eukaryotes. While never specifically codified, it is possible to articulate an implicit model of dinoflagellate genome organization that is widely shared and has shaped the understanding of dinoflagellate genomes. Although there is a diversity of opinions regarding many aspects of this model, we believe that the general interpretation we present here is widespread, and refer to it as a “consensus model.”

The consensus model suggests that: 1) dinoflagellate genes are highly duplicated and organized in tandem repeats, 2) genes of the tandem repeat (hereon referred to as tandem repeat members or “members” for short) are found in long arrays, encoding isoforms of the same protein, with synonymous substitutions at nearly every available site and rare amino acid substitutions, 3) Traditional eukaryotic promoter, terminator, and intron boundary sequences are thought to be absent, 4) introns are rare and tend to be small when present, 5) a 22 base pair (bp) sequence, encoded in a separate gene, is trans-spliced to the 5' end of pre-mRNA transcripts to form a mature mRNA.

Prior to 2007, discussion of this unusual feature set was, to our knowledge, always couched as applying only to the specific genes under discussion. A general consensus was reached when 3 publications in 2007 and 2008 all described these unusual features as being broadly

representative of dinoflagellate genomic organization [19,51,52]. Since then, the features of this consensus model have been found listed as general features of dinoflagellates in most publications on dinoflagellate biology. While no author has apparently felt sufficiently confident in the model to codify it, the model has nevertheless shaped discussion, analysis, and understanding of dinoflagellates for nearly a decade.

Dinoflagellate introns are also unusual. Intron splice-site consensus sequences in most eukaryotes conform to the consensus sequence MAG|GTRAGT at the 5' splice site and CAG|G at the 3' splice site [53,54]. The most common, non-canonical, splice-site consensus sequence uses GC at the 5' splice site rather than the canonical GT, but otherwise conforms well to the remaining consensus and is spliced by the same spliceosomal complex as canonical introns [55,56]. A rare class of introns, spliced by a separate spliceosomal complex, conforms to the consensus sequence RTATCCTY at the 5' splice site. These introns account for a small percentage of introns in a variety of eukaryotes including *Homo sapiens*, *Drosophila melanogaster*, and *Arabidopsis thaliana* [57,58]. Dinoflagellate introns, when observed, have been noted to not conform to any known splice-site consensus sequence, nor to have the secondary structure characteristic of self-splicing introns [15,17,19,21,59].

We report here the map-based Sanger sequence of a 39,500 bp cosmid containing three entire and two partial copies of genes encoding members of the Alcohol Dehydrogenase (ADH) superfamily, derived from the *C. cohnii* genome. Although labor-intensive, this approach has the advantage of requiring neither conserved PCR primer sites nor the assumptions of short-read sequence assembly. These data provide unique insights into genome structure of *C. cohnii*.

Materials and Methods

### 2.1.1 Nucleic Acid Isolation

Genomic DNA was isolated from *Cryptothecodinium cohnii* Seligo strain “KO”, an axenic, monoclonal isolate from the non-clonal culture ATCC #30340. The KO culture was grown to a concentration of approximately  $10^7$  cells/ml in a medium containing 50 g/l glucose, 6 g/l yeast extract, 32-ppt artificial seawater, pH 6.7 at 27 °C shaking at 200 rpm. Cells were harvested by centrifugation at 4 °C and 3,000 g for 20 minutes. Cell pellets were transferred to plastic bags and flash frozen in liquid nitrogen. The frozen pellets were ground to a fine powder with a liquid nitrogen cooled mortar and pestle. Thirty grams of frozen-powdered biomass was mixed with 100 ml extraction buffer (100 mM Tris, 1.5 M NaCl, 50 mM EDTA, 2% w/v Cetrimonium bromide, 50 mM dithiothreitol, 100 U RNase) and warmed to room temperature in a water bath. When the frozen pellet had thawed and was resuspended in extraction buffer, lysis was allowed to continue at room temperature for an additional 5 minutes. DNA was extracted twice with an equal volume of a phenol-chloroform-isoamyl-alcohol mixture (25:24:1, v:v:v). Residual phenol was then removed with a chloroform-isoamyl alcohol (24:1, v:v) solution. DNA was precipitated with 2 volumes of 95% ethanol and 0.3 M Sodium Acetate at -20 °C for 1 hour. DNA was pelleted by centrifugation at 3,000 g for 30 minutes at 4 °C, washed with 70% ethanol and resuspended in 10 mM Tris to a concentration of 1 µg/µl. Cells for RNA extractions were collected and ground using a mortar and pestle as previously described for DNA extraction. RNA Isolation from the frozen powdered pellet was performed using Ambion's RNAqueous Kit (Life Technologies, Grand Island, NY).

### 2.2 Cosmid Library Construction and Screening

The cosmid library was constructed according to Sambrook [60] using Agilent's SuperCos1 Cosmid Vector Kit (Agilent, Santa Clara, CA), XL1 Blue *E coli* cells, and Gigapack III XL

(Agilent, Santa Clara, CA) packaging kit. The library was plated, 60,000 cfu per plate, on Whatman Nytran N (Whatman, Maidstone, Kent, UK) filters layered atop an LB ampicilin plate. Replica filters were produced and allowed to grow overnight before being stored at 4 °C. Cell lysis and nucleic acid crosslinking was performed per Whatman's protocol, with the modification that cell debris was vigorously scraped off the filters in the 2X SSPE bath followed by a brief rinse in 2X SSPE. An alcohol dehydrogenase gene with a highly abundant transcript was selected from an existing EST library for isolation. Probes for screening the library were made using a previously isolated cDNA clone (GenBank Accession KJ831651) by restriction digest and radiolabeled using Promega Prime-a-Gene Labeling System (Promega, Madison, WI). Probes were hybridized to primary filters according to Sambrook [60] using Church Buffer in thermal-sealed plastic bags in a shaking water bath, and washed according to Sambrook [60]. Positive colonies were identified following overnight exposure of the filters to a phosphor imaging screen and imaged with a phosphorimager. Images produced by the phosphorimager were used to correlate positive signals to colonies on the replica plates. Putative positive colonies were picked and rescreened by the same process until pure colonies were isolated. Cosmids were isolated from 500 ml broth cultures of positive colonies using Qiagen Plasmid Maxi Kit (Qiagen, Venlo, Netherlands).

#### 2.2.1 DNA Sequencing and Analysis

The cosmid was sequenced by Eurofin's MWG Operon transposon-based sequencing service (Eurofin, Luxenberg). Analysis of the sequence was performed in Biomatters Ltd Geneious software package (Biomatters, Auckland, New Zealand). Alignments of the cosmid insert sequence to previously identified cDNAs were used to identify gene repeats, intergenic spacers, exons and introns. The open reading frame of each member, including putative start and stop

codons and splice-leader acceptor sites were identified by comparison to previously identified cDNAs and manual examinations of the alignments.

Analysis of the HCc gene previously published involved a BLAST search of a proprietary EST library using the HCc sequence as a query term and alignment of all matching ESTs to identify the putative reading frame and splice sites.

## 2.3 Results

### 2.3.1 Analysis of Cosmid Sequence

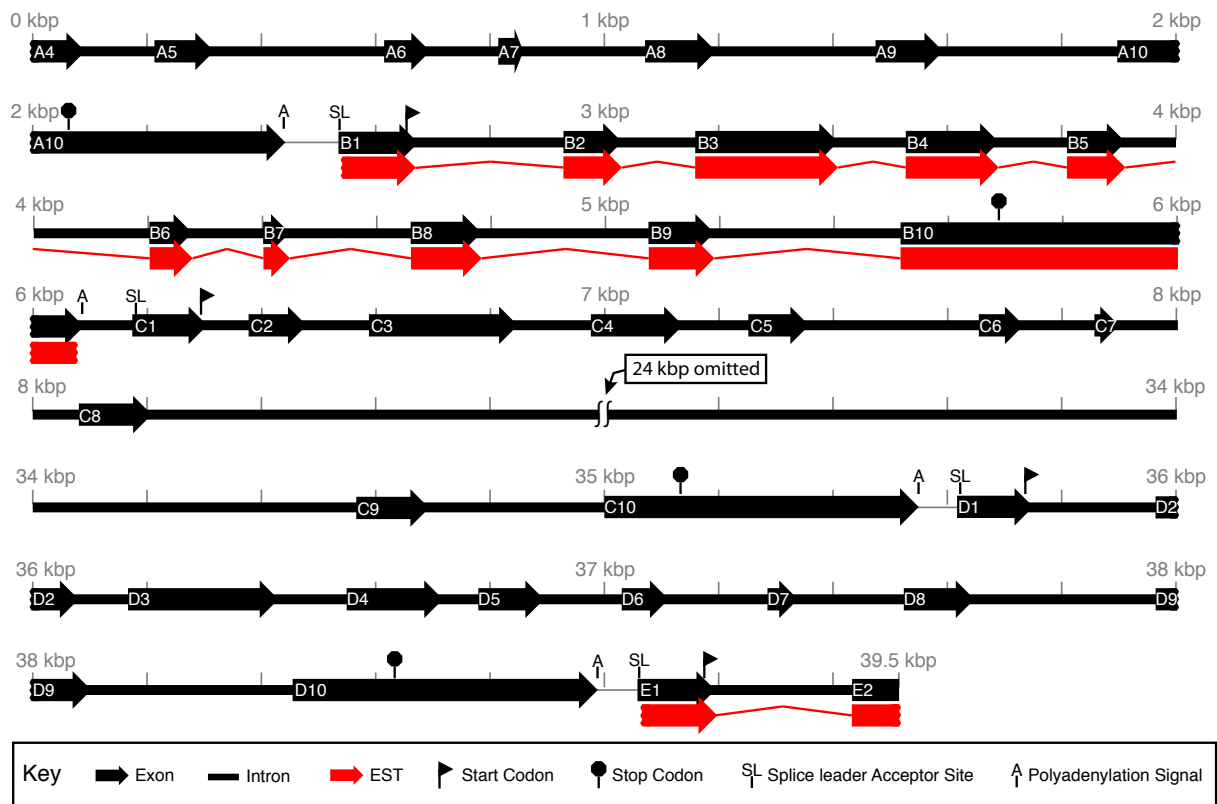
A *Crypthecodinium cohnii* cosmid library was screened using a probe for the gene ADH. One colony on nearly every initial plate that was screened, hybridized to the ADH probe (~1.5 in every 60,000 colony forming units). Upon re-screening, fewer than half of these signals were confirmed. Northern blots were also performed for ADH to establish that we were not observing a polycistronic mRNA (data not shown).

### 2.3.2 Map of Cosmid Sequence

Of several clones eliciting a confirmed ADH hybridization signal, one was selected for scale-up and full DNA sequence analysis. This cosmid (GenBank Accession KJ831652) was found to contain an insert of 39,500 bp. The probe sequence and other ADH sequences from the EST database were compared to the cosmid sequence to map gene boundaries. One region of the cosmid was found to be a perfect match in 10 discrete exons to a cDNA (GenBank Accession KJ831649) while the 3' end of the cosmid sequence clipped by the insert ligation point matched another cDNA (GenBank Accession KJ831650). Using this approach, a total of five similar but non-identical ADH gene copies could be annotated in the cosmid sequence, designated here A



through E (Figure 2). The center three copies appear to be complete, and two copies are truncated by the ligation points of the insert to the cosmid vector, one at each end of the insert.



**Figure 2 – Schematic of Alcohol Dehydrogenase Cosmid Insert**

Schematic of alcohol dehydrogenase (ADH) cosmid insert. The exons, poly-adenylation signals, splice leader acceptor sites, and start and stop codons are all indicated. The transcripts as they align to genomic sequences are indicated. Exactly 24,000 bp from the middle of the eighth intron of member ADH-C has been removed for illustrative purposes.

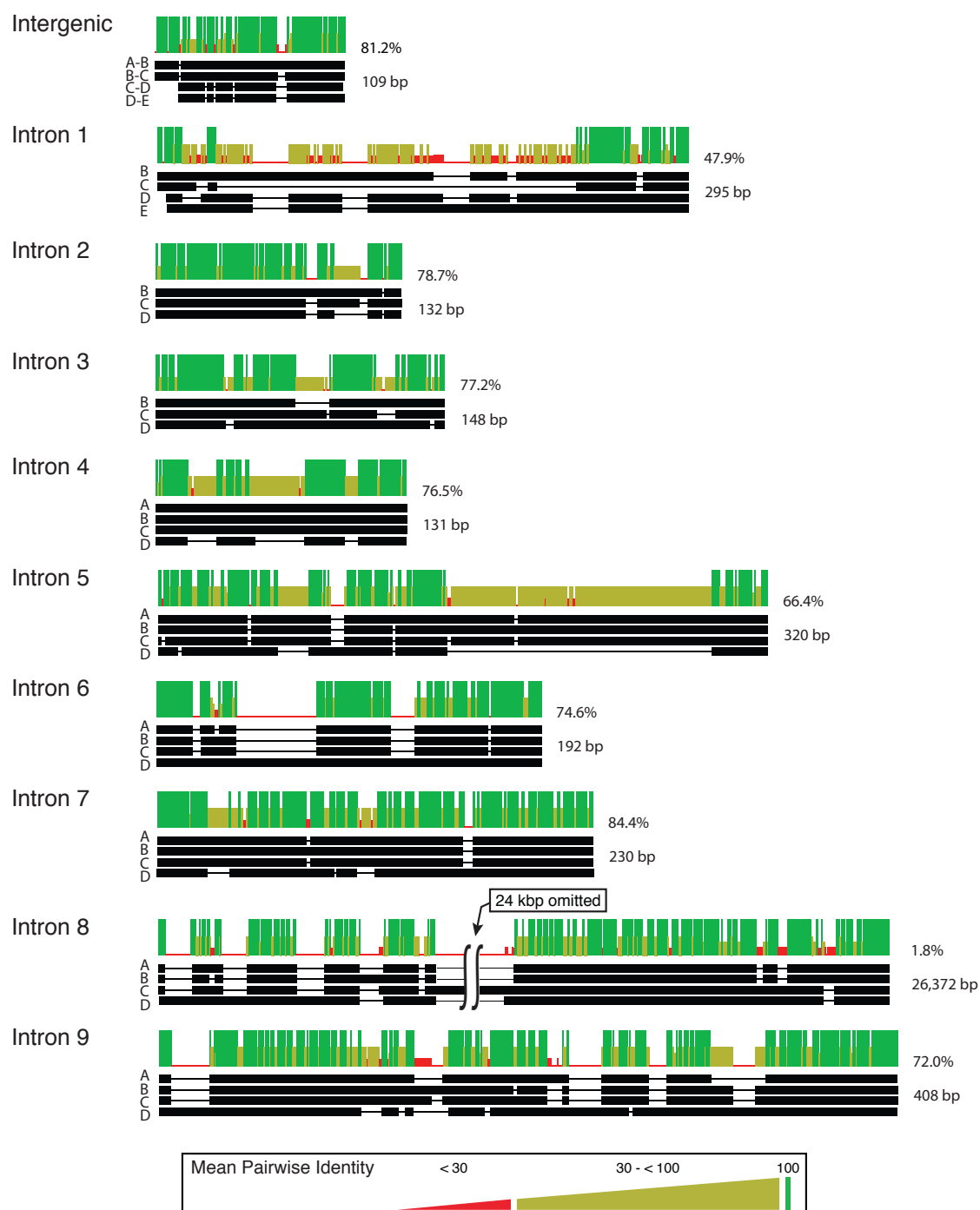
These initial predictions were modified to take into account the putative locations of splice-leader acceptor sites and poly-adenylation signal sites (Figure 2). The SL-acceptor site was identified as the first AG upstream of the aligned EST sequence. This change extended the 5' UTR of ADH-B and ADH-E 16bp and 14bp further upstream of the aligned ESTs KJ831649 and EST KJ831650 respectively. While a variety of potential alternative SL-acceptor sites can be observed in the 5' UTR of all the observed gene repeats, the first available AG upstream of the previously predicted 5' UTR was used, which corresponded to the same location in all the gene repeats. The poly-adenylation signal was marked as 5'-AAAAACAAAAA-3' or 5'-AAAAACAACAA-3'. This extends the 3'UTR of each gene 11bp past where the aligned EST begins a poly-adenylation sequence. Start and stop codons were identified by aligning all available ADH ESTs with the ADH gene repeats from the cosmid. A sharp drop-off in sequence conservation marked the extremities of the coding sequences and allowed start and stop codons to easily be identified.

### 2.3.3 Sequence Similarity Analysis

A total of 34 introns were identified, all but one of which ranged in length between 83 and 371 bp, with a median size of 209 bp. The exception was a large intron measuring 26,372 bp located between the seventh and eighth predicted exons of ADH-C. The coding regions of the complete exons ranged in size from 6 bp to 246 bp with a median size of 92 bp. The coding region of the first exon of each member was particularly small; just 6 bp. The coding regions of the exons comprise just 9.9% of the cosmid insert, and just 29.8% of the insert when the particularly large intron is removed from the calculation.

The coding regions of the members were highly conserved. The pairwise nucleotide identity of the exons in the 5 paralogs observed in the cosmid insert is 97.7%. The pairwise identity of the

amino-acid translation is 99.8%, differing in just one amino acid where the codon TTT, for Phenylalanine in ADH-C is TCT, for Serine, in the other members. Non-coding regions were less conserved, but are still very similar (Figure 3).



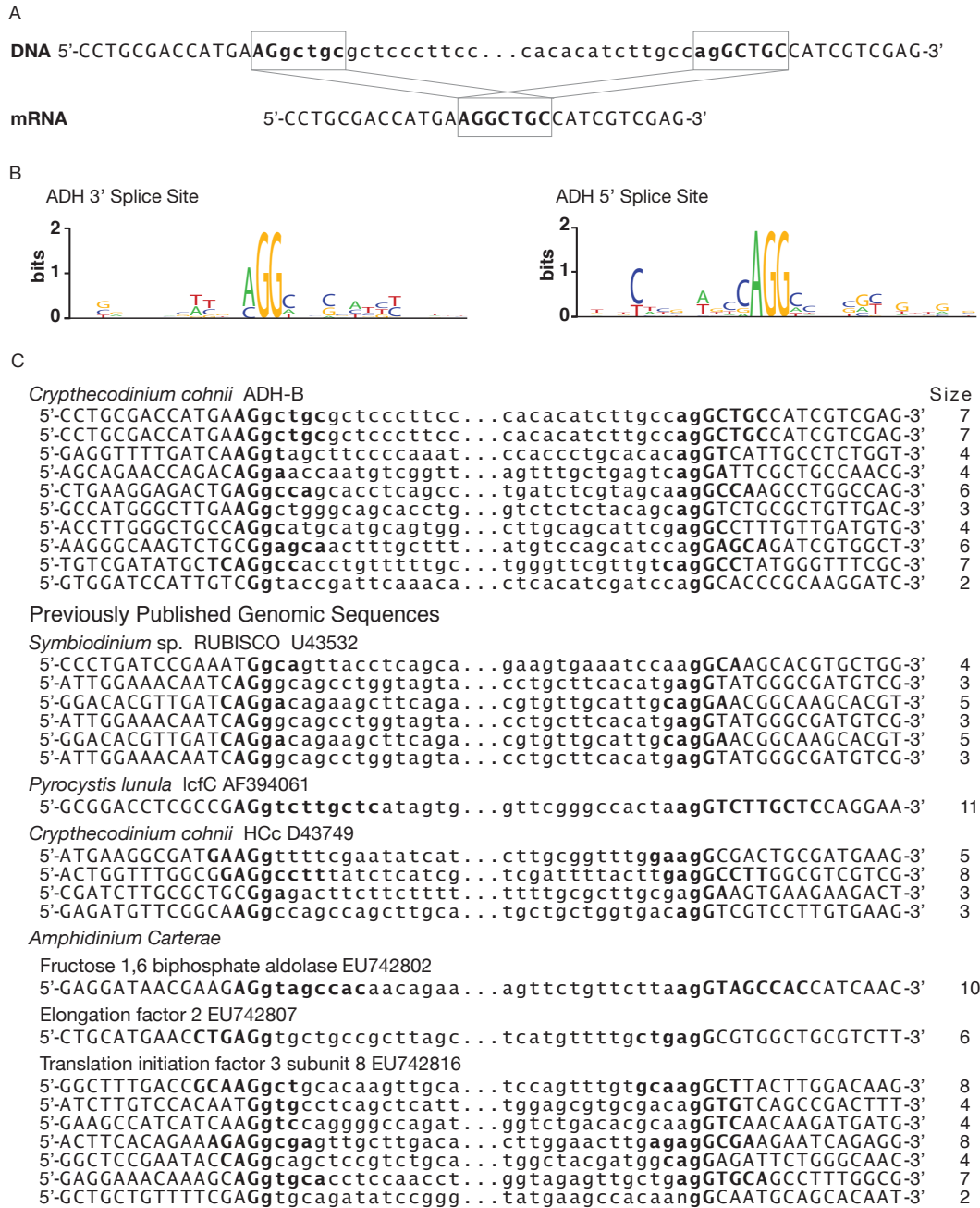
**Figure 3 – Pairwise Identity Graphs for Intergenic Spacers**

Alignments and mean pairwise identity graphs for the intergenic spacers between alcohol dehydrogenase (ADH) members and the nine introns of the five members of ADH. Length and pairwise identity is displayed to the right of each alignment. Gene sequences are indicated by black bars, and indels are indicated by intervening black lines. Each column of the alignment has a bar graph above it indicating the mean pairwise identity of all pairs in the alignment. The height of each bar in the graph is proportional to the mean pairwise identity of all pairs in that column and the color of the bar green at 100%, yellow between 30 and less than 100%, and red when less than 30%.

Indels in non-coding regions significantly lowered calculations of pairwise identity. The pairwise identities of the introns range from 1.8% to 84.4%. The fourth introns of ADH-A and ADH-B are identical and the sixth introns of ADH-B and ADH-C are identical. Alignment of intron 8 revealed the large intron from ADH-C has a 26,073 bp insertion compared to the other copies, accounting for 66.0% of the entire cosmid insert. The intergenic spacers range in size from 86 bp to 108 bp. The pairwise identity of the intergenic spacers is 81.0%, with 9.5% of the consensus sequence composed of gaps.

#### 2.3.4 Intron Border Repeat (IRIB)

The intron junctions of ADH-B were closely examined for potential splice site consensus sequences. A unique pattern was discovered in which the last 2 to 9 bp on the 3' end of each exon exactly matched the 3' end of the immediately downstream adjacent intron (Figure 4A).



**Figure 4 – Identical Repeated Intron Boundary**

Nucleotide sequences surrounding the intron splice sites of alcohol dehydrogenase (ADH) from *Cryptocodinium cohnii* as well as previously published genes of various dinoflagellates. **A.** Genomic and mRNA sequences of the intron splice sites of ADH-B intron 1 indicating the identical repeated intron boundary (IRIB) sequence is present at either end of the intron twice, but is present in the mRNA only once. **B.** Sequence logos, generated by WebLogo, using the intron splice sites from all introns present in the ADH cosmid sequence. **C.** Intron splice sites of ADH-B and previously published dinoflagellate genomic sequences indicating the IRIB sequence and the size of the IRIB. Each IRIB sequence is indicated in bold and their corresponding sizes are listed at right. Portions of each internal intron sequence have been replaced by ellipses as shown.

This identical repeated intron boundary (IRIB) sequence is found only once in the corresponding cDNA, therefore one could annotate the sequence so that one or the other IRIB is annotated as being exonic or intronic, or even that a part of each IRIB contributes to the translated sequence (Figure 4A). Here the splice-sites are always placed between the conserved GG as found in canonical U2 splice-site consensus sequences. When the exons are defined in this manner, the exons from each member fall in the same locations. If a different convention is used, the exons may differ in length by several base pairs. The repeat always contains a GG and often an AGG at the 3' splice site and always has an AGG at the 5' splice site (Figure 4B). The rest of the repeated sequence was unique to each splice-site.

Evidence of IRIB sequences was sought in previously published data from other researchers. RUBISCO from *Symbiodinium* sp. and LCF from *Pyrocystis* both show IRIB sequences in their published introns as annotated by their authors (Figure 4C). The introns from the gene HCc from *C. cohnii* do not show an IRIB sequence as published, however the intron exon boundaries annotated by the authors could not be established with confidence since the authors lacked a mRNA sequence with 100 percent identity. Using the HCc gene as a query sequence against our own *C. cohnii* EST library revealed several ESTs that allowed confident re-annotation of the exon/intron boundaries of the previously published HCc genomic sequence and revealed IRIB sequences at every intron (Figure 4C). Analysis of genes from the *Amphidinium carterae* survey revealed many introns with IRIB sequences, a subset of which are pictured in Figure 4C.

## 2.4 Discussion

### 2.4.1 The Consensus Model

To the best of our knowledge, other than the *Symbiodinium* genome survey, these data represent the longest contiguous dinoflagellate genomic tandem array yet published, and has the advantage

of being sequentially sequenced. While dinoflagellate genes are known to be present in tandem repeats and it has been inferred that many copies exist in tandem arrays, this sequence is unique in containing three complete gene duplicates and two more flanking gene duplicates for a total of five tandem genes. Previous evidence of multiple genes in tandem from dinoflagellates included very small genes, in the case of the gene encoding the splice-leader, alignments that accept a small amount of mismatch in overlapping sequence and thus do not necessarily represent sequences that were physically adjacent to each other, or assembly from very short reads [15,16,18,19,32,44-48,50,59,61]. This longer contiguous copy set provides additional evidence that genes arranged in a common array all encode the same protein, consistent with the consensus model.

Some evidence suggests that the consensus model does not best describe all dinoflagellate genes [19]. There may in fact be two models of genes that are organized and transcribed differently from one another [19]. The first model is typified by genes that are organized in tandem repeats, present in high copy number, highly expressed, trans-spliced with a conserved leader sequence, and have low intron density [19]. These genes can be considered the consensus model group. The second model of genes are not well studied, but seem to be organized like classic eukaryotic genes [19]. These genes may have eluded initial detection because they are found in low copy number and are transcribed at much lower levels than the genes in tandem arrays. These genes are intron-rich, are not trans-spliced, are transcribed at low levels, and contain common eukaryotic motifs for transcription and RNA processing [19]. Testing the comprehensiveness and accuracy of the consensus model is beyond the scope of this study, but it is important to note that our data are difficult to reconcile with either of these models. The consensus model was shaped largely by data collected using PCR: of the 11 gene sequences that our research indicates



have contributed to the consensus model, 10 have been isolated using PCR methods [15,16,18,32,44-48,50,59]. The consensus model also seems to be in conflict with data from the *Symbiodinium minutum* genome survey. Two major disagreements between the consensus model and the *Symbiodinium minutum* genome survey are the paucity of genes arranged in tandem and the high frequency of introns in the *Symbiodinium minutum* data. The data we present here deviate from the consensus model principally in the high frequency of introns. Whether the organization of this gene cluster is representative of the rest of the *Cryptothecodinium cohnii* genome and whether *C. cohnii*'s genome is broadly representative of dinoflagellate genomes is unknown, but it is notable that the kind of biases expected from a model developed using PCR data, happen to be the very areas that conflict with sequences collected via cosmid library screening. Selection of genes that are short enough to amplify by PCR could have resulted in a model built upon a non-representative set which lack introns or have unusually few and small introns. How the *Symbiodinium minutum* genome survey fits in is unclear. The scarcity of genes in tandem could be the result of incomplete assembly and gene modelling or could be real and simply the result of an endosymbiotic lifestyle. Whether the *Symbiodinium minutum* genome is representative of broader dinoflagellate genomic organization or not, it highlights the need to vigorously test the consensus model with broader data sets.

#### 2.4.2 Conservation of Non-Coding Regions

The observation that dinoflagellate genomes are often organized into tandem gene arrays has led to speculation on the evolutionary processes underlying this organization [62]. The presence of what appear to be vestigial splice-leader sequences in several dinoflagellate transcriptomes led to the inference that dinoflagellate genes could be duplicated via reverse transcription and reintegration into the genome of mature, trans-spliced transcripts [62]. Because introns would

regularly be purged via such reintegration of reverse-transcribed mature transcripts, this hypothesis predicts that introns would be rare, and when present would have been inserted relatively recently. Our observations conflict with that hypothesis, providing evidence for a gene-duplication mechanism that preserves intron/exon structure. Assuming our splicing inferences are accurate, the relative intron positions of all five ADH members are perfectly conserved. Furthermore, the sequences of corresponding introns are also conserved. These observations are inconsistent with an mRNA intermediary and reintroduction of introns after duplication. Nor was there any evidence of vestigial splice-leader sequences in any of the ADH members. If reverse transcription does play a role in the duplication of dinoflagellate genes, it is unlikely to have been the process that created the gene cluster described here. The conservation of intron splice sites and sequences suggests a genome-level duplication mechanism, as well as either relatively recent duplication or concerted evolution (or both).

Whether the gene duplication of members of a tandem array in dinoflagellates has arisen due to concerted evolution or whether it represents a birth-death model has been examined in some detail for both actin and Peridinin Chlorophyll-a binding protein genes of *Amphidinium carterae* and *Symbiodinium* respectively [31,48]. In most eukaryotes, the sequence uniformity in tandem arrays of rRNA genes is thought to be maintained by concerted evolution. In concerted evolution uneven crossing over and gene conversion result in high sequence similarity between members of an array. In a birth-death model, sequence similarity is maintained via purifying selection of the encoded proteins. In an array functioning under a birth-death model only non-synonymous substitutions will be homogenized. Since the process underlying concerted evolution affects synonymous and non-synonymous substitutions equally, a comparison of the number of synonymous substitutions to non-synonymous substitutions between members of an array can

reveal the dominant contributing model. Analysis of members of a PCP tandem array led Reichman et al. to conclude that similarity was maintained via low-levels of concerted evolution. The analysis of 142 members of actin by Kim et al., however, indicated that a birth-death model best explained the similarity of members. Our data are consistent with the findings of Kim et al. of the birth-death model. Duplication events via uneven crossing-over and gene conversion cannot account for the differences in sequence similarity between coding and non-coding regions and the prevalence of synonymous substitutions in the coding regions. The birth-death model of gene duplication best explains the observed similarity of tandem array members, where the majority of changes occur in regions that do not affect protein structure. The similarities of non-coding regions are, however, still striking. It is possible that while most of the similarity between array members is maintained by purifying selection, low levels of concerted evolution are still at work.

#### 2.4.3 Non-Canonical Splicing of Introns

Intron splice sites in eukaryotes consist of a CAG|G at the 3' acceptor site and MAG|GTRAGT at the 5' donor site. While few dinoflagellate genes with introns have been sequenced, the unusual lack of the canonical GT-AG consensus sequencing denoting intron splice sites in dinoflagellates has been noted in every case [15,17,19,21,49]. Two of these authors noted a repeat at the ends of introns, but did not fully describe the pattern [15,19]. Within our data there is a consistent splicing pattern that is also consistent with most other published splice sites from dinoflagellates. The AG|G of the 3' and 5' splice site is usually conserved, and the splice donor and acceptor sites have a duplicate 2-11 bp sequence flanking the intron which remains in the mature mRNA only once (Fig. 2). Consequently, this creates ambiguity in the exact annotation of splice sites anywhere within the IRIB. We believe that the splice site we have designated here

is correct and consistent with other studies, but this ambiguity has resulted in understandable variation in the annotation of exact intron boundaries in other studies. Intron splice sites in HCC from *C. cohi*, lfcC from *Pyrocystis lunula*, and sequences from the survey of *Amphidinium carter* were all consistent with splicing as inferred here, although annotated slightly differently [15,19,21,59]. Analysis of previously published dinoflagellate genomic sequences containing introns reveals that the IRIB sequence is present in the all dinoflagellates for which there are available data, but the inherent flexibility of IRIB annotation makes the pattern difficult to recognize [15,17,19,21,49]. Interestingly, the splice site logo generated in the *Symbiodinium minutum* genome survey looks very similar to the one generated here. The major difference between the *Symbiodinium minutum* logo and our *C. cohnii* ADH logo is in the nature of the GG at the 5' donor site. This GG is conserved in *C. cohnii* ADH sequences and all other published dinoflagellate introns, but is not well conserved in the splice site logo generated for the *Symbiodinium minutum* genome survey. With the continued improvement of sequencing technology, more dinoflagellate genomic data is surely forthcoming; hopefully discovery of the dinoflagellate IRIB will improve the automated gene modeling necessary in such large scale sequencing projects.

### 3 A Method of Screening Genomic and Transcriptomic Libraries for Probable Orthologs: A Stramenopile Case Study.

#### 3.1 *Background*

Ortholog identification is an important aspect of modern biological science. Software such as the Basic Local Alignment Search Tool (BLAST) and HMMER, as well as more sophisticated tools built upon them, such as Core Eukaryotic Gene Mapping Approach (CEGMA), Benchmarking Universal Single-Copy Orthologs (BUSCO), and OrthoMCL, infer orthology based upon sequence similarity as a function of evolutionary origin[63-66].

Advances in high-throughput sequencing methods continue to accelerate the rate at which genomic and transcriptomic scale data becomes available to researchers. Following assembly of one of these deeply sequenced data sets, a common first step is to measure the completeness of the assembly using tools based upon ortholog searches. While a variety of tools exist for these purposes, they typically evaluate each potential ortholog in isolation[64,65,67,68]. Phylogenetic pipelines of increasing complexity are being made available[69,70]. While these pipelines make use of the widespread availability of genomic and transcriptomic data sets, they continue to be anchored purely in similarity-based ortholog identification in which each putative ortholog is considered in isolation.

Here I implement a ortholog identification and phylogenetic pipeline that builds upon the approach of CEGMA, but also attempts to leverage genomic and transcriptomic wide data sets to inform and filter the initial ortholog identification using phylogenetic approaches where each putative ortholog can be evaluated in concert with the others being investigated. In this manner

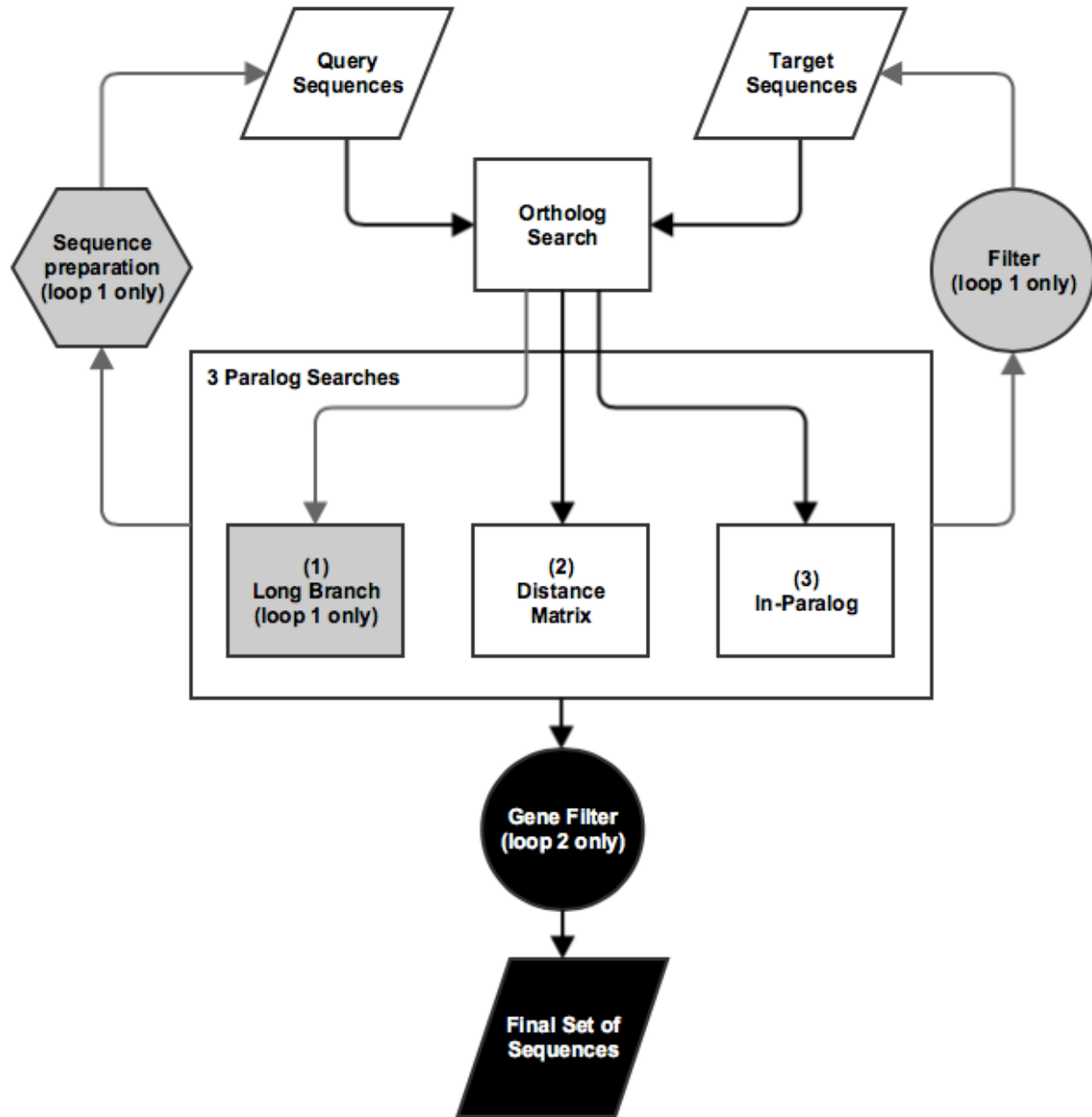
both the depth of the sequencing and the breadth of the taxon sampling provides useful information in the identification of orthologs.

## 3.2 *Implementation*

### 3.2.1 Overview

The pipeline begins with an initial search based on the approaches taken by the CEGMA and BUSCO pipelines, where BLAST search hits are ranked according to the bitscore of an HMMsearch performed using a Hidden Markov Model (HMM) built from the available query sequences for the given gene[63-65](see Chapter 3.2.2).

The presence or absence of each queried gene for each Operational Taxonomic Unit (OTU) is presented to the user with instructions to select a subset of genes and OTUs that have few data gaps to move forward in the pipeline. Data gaps are particularly problematic in this type of analysis, since the selection and screening of each sequence is improved by statistics garnered from the totality of the data. In order to generate new query files and to filter the target sequences of non-orthologous sequences, individual gene trees, concatenated alignment trees, and distance matrices are used in three tests: (1) an outlier analysis of branch-lengths of gene trees calculated using the concatenated alignment tree as a constraint tree (see Chapter 3.2.3), (2) an outlier analysis of the pairwise distances between each pair of OTUs in a sequence alignment to the same pairwise distance in all other sequence alignments in the data-set (see Chapter 3.2.4), (3) an analysis of gene trees with non-top hits included in the tree to identify and remove putative in-paralogs as well as putatively non-orthologous sequences (see Chapter 3.2.5)(Figure 5).



**Figure 5 – Overview of Pipeline**

Two rounds of ortholog searches are performed. The first round of searches creates a preliminary set of sequences that are used in three types of paralog analysis. These analyses are used to produce a new set of query terms and to filter the initial target sequences. A second round of searches is then performed using these new query sequences and filtered target sequences. Following the second round of searches, the genes are filtered to produce a set of genes where the pipeline was able to separate the paralogs from the orthologs. Portions of the flowchart only performed in the first part of the pipeline are shaded grey, while portions of the flowchart only performed in the second part of the pipeline are in black.

The newly generated query files and filtered target sequences are then used in a second round of searches. This second set of sequences is then evaluated a final time using the in-paralog and distance-matrix methods previously mentioned. Here the candidate gene clusters are filtered to produce a revised set of genes in which the pipeline was able to separate putatively orthologous sequences from non-orthologous sequences (Figure 5). The purpose of this round of tests is to identify entire genes that may be unsuitable for use in the final tree finding. In this way, genes which the pipeline has not been able to reliably separate putatively orthologous sequences from non-orthologous sequences will not be used for phylogenetic purposes. A final set of highly screened genes and putative orthologous sequences is now available for tree finding. A Hyper Text Markup Language (HTML) report is also available providing detailed tables and figures for each gene indicating to the user how each sequence was evaluated at each step of the pipeline and where along the pipeline various sequences were flagged as non-orthologous or passed all tests to be marked as a putative ortholog.

A case study using publicly available transcriptomic and genomic data sets from Stramenopiles is presented as an example of an analysis this pipeline can perform.

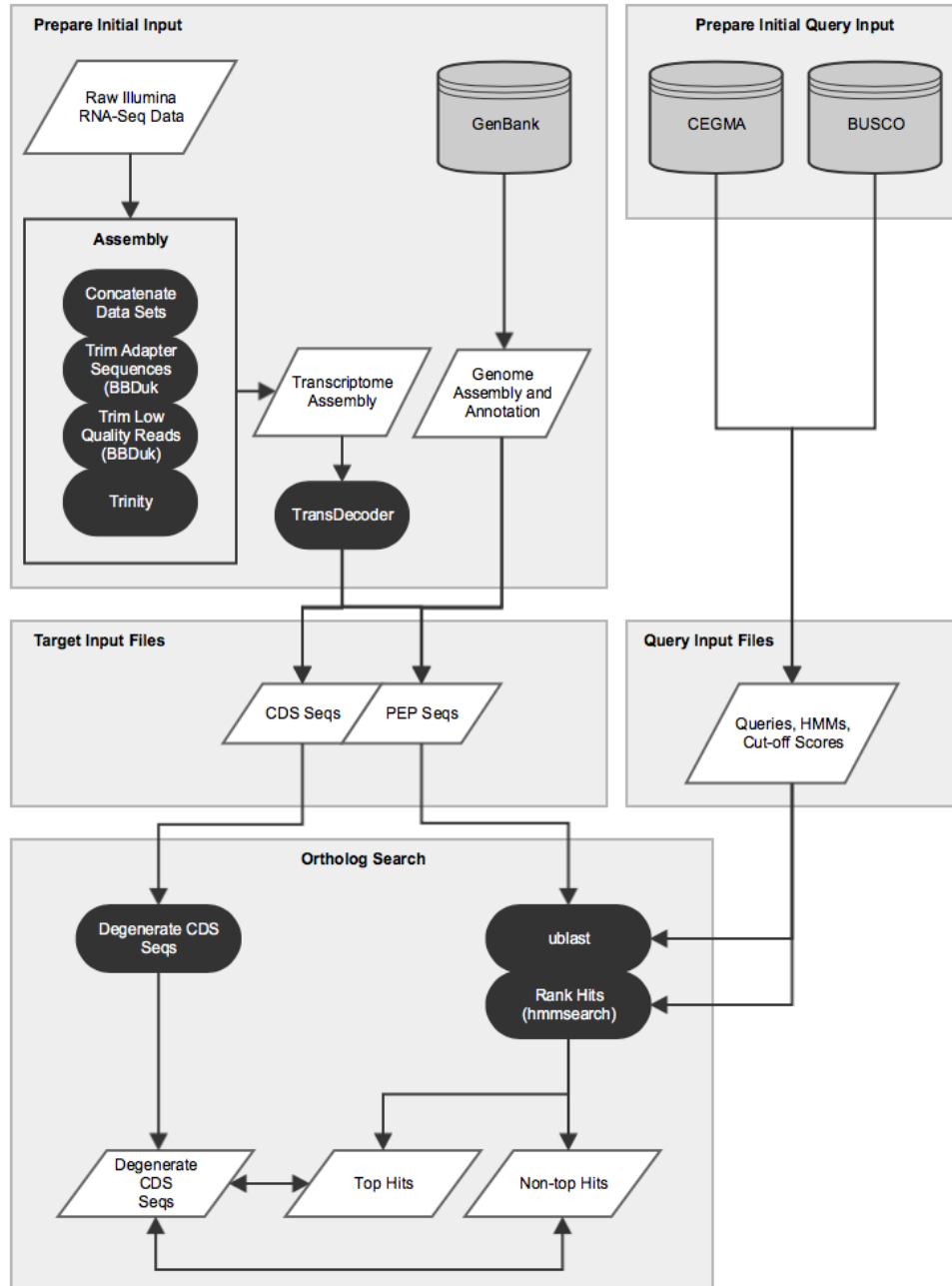
### 3.2.2 Ortholog Search and Scoring

The searches follow the same approach as the CEGMA and BUSCO pipelines[64,65]. This begins with a curated set of genes to use as query terms (Figure 6). A non-redundant set of 729 genes, HMMs, and HMMsearch bitscore cut-offs generated by the CEGMA and BUSCO projects is used as the initial query terms[64,65]. This set of starting genes can easily be expanded by the user by supplying FASTA files of the genes of interest to a set of auxiliary scripts that will produce the necessary HMMs and HMMsearch bitscore cut-offs. Each OTU of interest must have two associated FASTA files: one with amino acid sequences and one with the



corresponding nucleic acid sequences, with matching headers. Prior to the initial ortholog searches the nucleic acid sequences are degenerated to avoid GC biases in third codon positions from influencing the phylogenetic analyses. This degenerate version of the nucleic acid files will be used in all subsequent steps of the pipeline. Also prior to the initial searches, the amino acid files have duplicate and short sequences removed using usearch's --derep\_prefix option with a -minseqlength setting of 40[71]. This filtered amino acid FASTA file will be used in subsequent steps of the pipeline. Usearch is used with the -ublast option using the FASTA files of the curated genes as a query and usearch databases generated from the filtered protein sequence FASTA files as targets[71]. So each gene of interest is queried using a set of curated sequences.

Hits from usearch are ranked as described in the CEGMA and BUSCO pipelines, using HMMsearch, HMMs built from the query sequences, and cut-off scores[64,65,72]. Cut-off scores were calculated in the same manner as described in the CEGMA pipeline[65]. Hmmsearch bitscores for every query sequence against an HMM generated with all query terms, minus that query sequence being tested, are generated for each gene. These hmmsearch bitscores were averaged and divided by two to generate a hmmsearch bitscore cut-off for that gene. Only hmmsearch hits with bitscores above the cut-off score are recorded, multiple hits are ranked by bitscore, and each non-top hit must have a bitscore of at least 80% of the top hit's bitscore.



**Figure 6 – Flow Chart of Initial Ortholog Searches**

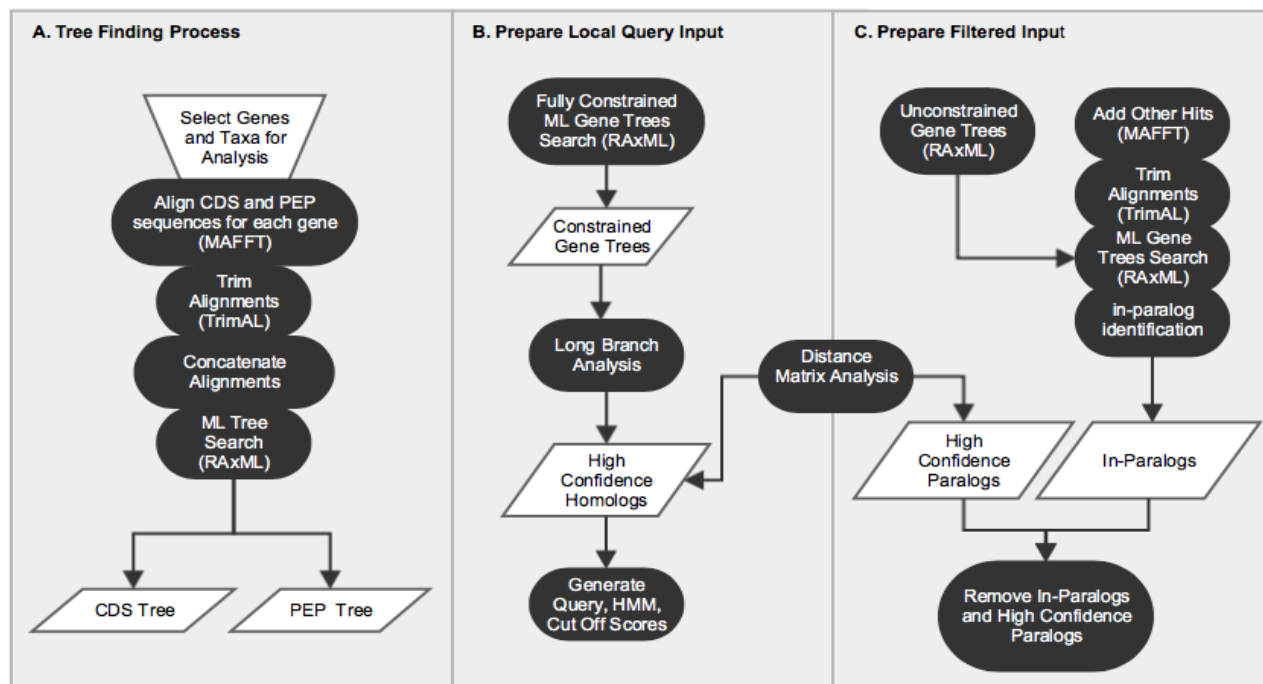
Initial searches are prepared in three steps. In this diagram input/output is denoted in white parallelograms, while actions are drawn in black rounded rectangles. First protein sequences and corresponding nucleic acid coding regions are identified for each transcriptome and genome to be queried. A set of query sequences for each gene of interest is used; in this case a non-redundant set of the genes identified by the CEGMA and BUSCO projects are used. uBLAST searches of all query files against all protein target databases produce an initial set of hits. HMMs for each set of query terms are then used in an HMMsearch to rank the uBLAST hits by HMMsearch bitscore. Each uBLAST hit must be equal to or greater than the cutoff score previously generated for the given gene. If multiple hits are above this cutoff score then the sequence with the highest bitscore is counted as a top hit and other hits must have bitscores of at least 80% of the top hit to be considered a non-top hit. When a set of amino acid sequences has been selected, the corresponding nucleic acid sequences are also selected for parallel analysis later in the pipeline.

Top and non-top hits are written to separate files and a comma-separated file is generated indicating to the user with ones and zeros which OTUs had at least one hit for each gene. The user is prompted to review this table and generate a list of OTUs and genes that will be allowed to continue downstream in the pipeline.

### 3.2.3 Constraint Tree Branch Length Outlier Analysis

The purpose of this analysis is to identify individual sequences that are incongruent with a concatenated alignment tree. The analysis is performed separately on both nucleic acid and amino acid data. A sequence that fails either analysis will not be eligible for use as a query sequence for the second round of searches. The cutoff to be considered incongruent should be set to low stringency. Sequences that pass this test will be used as query terms in a second round of uBLAST searches, so it is better to exclude marginal sequences than risk non-orthologous sequences being used as query terms.

The analysis begins following user input of selected OTUs and genes based on the coverage table supplied by the initial searches (Figure 7a).



**Figure 7 – Flow Chart of Paralog Filtering Process**

In this diagram input/output is denoted in white parallelograms, while actions are drawn in black rounded rectangles. The paralog filtering process involves three main steps. A) Multiple sequence alignments are prepared and concatenated trees are generated. B) A set of sequences for use as query terms for a second round of searches is produced by filtering out incongruent sequences identified by the long branch analysis and distance matrix analysis. C) Incongruent sequences identified by the distance matrix analysis and in-paralogs and sister paralog analysis are removed from the target sequence databases that will be used in a second round of searches.

New flat files are generated for each selected gene containing only OTUs selected by the user for both nucleic acid and amino acid data. Each data-set will proceed through the following analysis steps separately. Multiple sequence alignments (MSAs) for each gene are created using MAFFT v7.215[73]. MSAs are trimmed of poorly aligned sections using trimAl v1.2 [74] using the “automated1” algorithm. Trimmed MSAs for each gene are concatenated into a single MSA. The maximum likelihood tree is found using RAxML v8.1.24 [75]. Fully constrained individual gene trees are now calculated using a tree based on the concatenated alignment tree, differing only in that OTUs present in the concatenated MSA that are not available for each individual gene MSA

have been pruned from the constraint tree (Figure 7b). By fully constraining the tree, the RAxML software is restricted to calculating branch lengths of the given MSA and constraint tree.

Each constrained gene tree is analyzed separately to identify branch lengths that are anomalously long given the branch lengths that make up that tree. Unusually long branches are indicative of sequences that are highly incongruent with the given constraint tree, which could be due to contamination, horizontal gene transfer (HGT), or a paralogous sequence. To determine what constitutes an unusually long branch for a given set of branch lengths, an outlier analysis is performed using the Median Absolute Deviation (MAD):

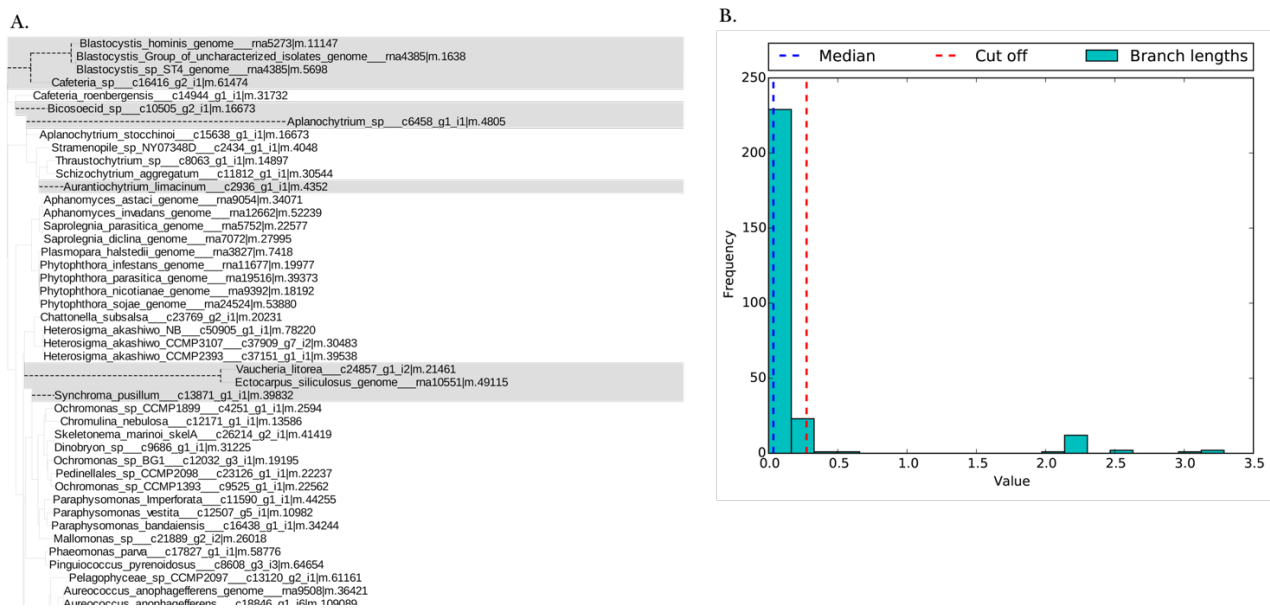
$$MAD = median(|x_i - median(x_i)|)[76]$$

A branch length is determined to be anomalously long in this analysis if the length is greater than a cut-off score equal to the median of all branch-lengths plus a multiplier of the MAD set by the user:

$$Cutoff = median(x_i) + Y * MAD(x_i)$$

This multiplier variable (Y) is exposed to the user during script execution, with a default setting of seven. Therefore, in the default setting, a branch length must be greater than seven MADs above the median to be flagged as anomalously long. All leaves of an anomalous branch are counted as incongruent sequences. To avoid potentially using non-orthologous sequences in the second round of searches, this cut-off score has been set to a low stringency. Some true orthologs will likely be flagged as incongruent in an effort to build a set of query terms for which we have the highest confidence are orthologous. These hypothetical orthologous sequences will likely still be recovered during the second round of searches.

The analysis produces a list of sequences for each gene analyzed that are incongruent with the given tree. These sequences are denoted as having failed the test in the HTML report. These sequences will not be used in the creation of new query terms, HMMs, and HMM bitscore cut-offs used in the second round of searches. The analysis also produces a histogram of branch lengths with the cut-off score indicated (Figure 8a), and a gene tree with flagged sequences highlighted (Figure 8b) for each gene and linked from the HTML report.



**Figure 8 – Histogram and Cladogram of Long Branch Analysis**

The long branch analysis examines the branch lengths of each gene tree to flag sequences with anomalously long branches relative to the overall branch lengths in the given tree. (A) An example of the long branch analysis of the gene BUSCO 004338, branches indicated with grey backgrounds and dotted branches have been flagged as anomalously long. (B) The corresponding histogram of the gene tree branch lengths indicating in blue the median branch length and in red the cut-off branch length. Note that only a portion of the entire gene tree is pictured.

### 3.2.4 Distance-Matrix Outlier Analysis

To complement the previous test, which had low stringency, a second test is performed that will flag sequences with higher stringency. In this test pairwise distance matrices for each gene alignment are analyzed to identify anomalous pairwise distances, and thereby flag individual sequences (Figure 7b-c).

The previously generated trimmed MSAs for each gene are used by RAxML to generate pairwise distance matrices for each gene. Before pairwise distances between different genes can be compared the data must first be standardized. Distances are logarithmically normalized, and standardized using the median pairwise distance across orthologs to produce a standardized normalized distance score  $D_{ab}$  defined as:

$$D_{ab} = \ln\left(\frac{S_{ab}}{\widetilde{S}_{ab}}\right)$$

where  $S_{ab}$  is the individual pairwise distances and  $\widetilde{S}_{ab}$  is the median of all pairwise distances across putative orthologs. These normalized-standardized pairwise distances are produced similarly to methods used in alignment-free phylogenies and rate-based methods of identifying selective signatures [77,78].

A modified Z-score, defined as:

$$M_i = \frac{0.6745(x_i - \text{median}(x_i))}{MAD} \quad [76]$$

with MAD denoting the median absolute deviation defined as:

$$MAD = \text{median}(|x_i - \text{median}(x_i)|) \quad [76]$$

is then calculated as part of a standard outlier analysis on the transformed distances to identify pairwise distances that may include a non-orthologous sequence. A modified Z-score was used instead of a Z-score, as is the best practice when dealing with data containing a large number of expected outliers which will exert a greater influence on a mean than a median[76]. A modified Z-score greater than 3.5 is the standard threshold to be considered an outlier[76]. Since this only exposes pairwise distances that are outliers rather than determining which sequence is anomalous, a percentage score of outliers is calculated for each sequence. A cut-off percentage score can be configured by the user, but defaults to 25%. Therefore, if a sequence has more than

25% of its pairwise modified Z-scores flagged as outliers, the sequence will be flagged as non-orthologous. A tree image is produced for each gene, highlighting sequences flagged as outliers (Figure 9) and linked from the HTML report.



**Figure 9 – Example of Distance Matrix Outlier Analysis Gene Tree**

In the distance matrix analysis, distance matrices for all genes are analyzed to identify sequences with a large number of anomalous pairwise distances. In this example of the gene BUSCO 004338, the anomalous sequences are displayed on a gene tree. Sequences with grey backgrounds and dotted branches have been identified as non-orthologs by the distance matrix analysis. The numbers to the right of each flagged sequence are the percentage of pairwise distances for the given sequence that have modified z-scores greater than 3.5. Note that only a portion of the entire gene tree is pictured.

This threshold is set to be highly stringent, only sequences with a large number of anomalous pairwise distance scores are flagged. Some non-orthologous sequences might not be flagged. The flagged sequences will not be used to generate query terms, and will also be removed from the target databases used in a second round of searches (Figure 7b-c). Since flagged sequences are



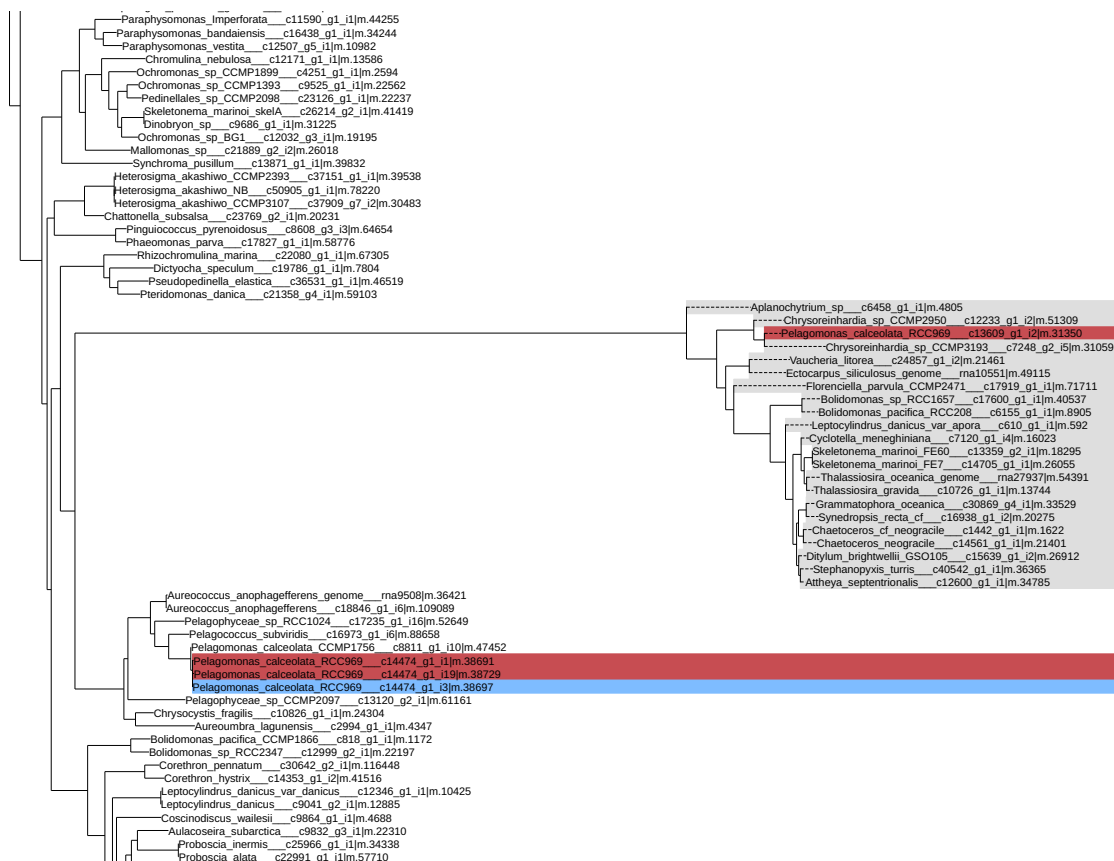
removed from the further analysis and consideration, only those sequences with strong evidence of non-orthology should be flagged.

### 3.2.5 In-paralog and Sister Paralog Search

In the previous two tests only the top hits, as ranked by HMMsearch bitscore, were tested. In this test, the non-top hits are also evaluated. The purpose of this test is to filter the target sequence databases, to be used in a second round of searches, of sequences that are in-paralogs or sister to the high confidence non-orthologous sequences flagged by the previously discussed high stringency distance matrix outlier analysis. This step is crucial to the final analysis that will evaluate the proper functioning of the pipeline for each gene under consideration. This final analysis makes use of statistics including the average number of total hits per OTU.

The previously generated trimmed MSAs are used to find gene trees using RAxML. For each gene, each species with multiple hits individually has the additional sequences added to the previously generated MSA using the `--add` feature of MAFFT. These new MSAs are then used to find new gene trees using the previously generated gene trees as a constraint tree to determine the placement of the new sequences in the gene tree. Each of these new gene trees is analyzed to identify in-paralogs. The sequence with the shortest branch length from a set of in-paralogs is retained and the longer branching sequences are flagged for removal from the target database. Remaining sequences are then tested to see if they are sister to a non-orthologous sequence identified by the high stringency distance matrix outlier analysis (see Chapter 3.2.4). Sequences sister to a putatively non-orthologous sequence are also flagged for removal from the target database. Gene tree images for each gene and each species with non-top hits are produced indicating which sequences were checked, and which sequences were flagged by this test as well

as the distance-matrix outlier analysis (Figure 10). These tree images are linked from the HTML report.



**Figure 10 – Example In-Paralog and Sister-Paralog Analysis**

In the in-paralog and sister-paralog analysis non-top hits are checked one species at a time. In this example from the gene BUSCO 004338 for the species *Pelagomonas\_calceolata\_RCC969*, four sequences are evaluated (in red and blue). Sequences flagged by the distance matrix outlier analysis are marked in grey. Sequences checked are indicated in blue, unless flagged for removal, then they are flagged in red. Sequences are flagged for removal if they are in-paralogs or if they are sister to sequences marked by the distance matrix analysis as non-orthologs. The sequence with the shortest branch length from a set of in-paralogs is retained. In this example one sequence was marked as paralogous since it was sister to sequences flagged by the distance matrix analysis. Two more sequences were flagged as being in-paralogs. In the end only a single sequence from this species will be retained. Note that only a portion of the entire gene tree is pictured.

### 3.2.6 Preparing Inputs for Second Round of Sequence Searches

Prior to starting a second round of searches, the output from the previously described tests is used to generate both new query terms and new filtered target databases. New target databases are generated by removing sequences identified by the distance matrix analysis and the

in-paralog and sister paralog analysis, from the original target databases. Sequences identified solely by the long branch outlier analysis will not be removed from target databases, but will not be used as query terms. Query terms should represent the set of sequences with the highest confidence of being orthologs. Only sequences passing all three tests are gathered into FASTA files for each gene, and HMMs and cut-off scores are calculated as previously described (see Chapter 3.2.2). The user can then execute a second round of searches that uses the newly generated query terms and filtered databases as input instead of the inputs used for the first round of searches.

### 3.2.7 Tests of Second Round of Searches

Following the second round of searches the user is presented with a comma-separated file with a new coverage table and a command that will initiate a final series of tests to determine which genes should be used for the final phylogenetic tree finding. The central questions addressed in these tests is whether the pipeline can identify a single putative ortholog for each OTU. If the pipeline is unable to find sequences for many of the taxa, finds multiple sequences for each OTU, or finds sequences that are still flagged as non-orthologous by the high stringency distance matrix and in-paralog analyses then the hits for the given gene should not be considered orthologous and should not be used as phylogenetic markers.

The checks start by running the distance matrix outlier analysis and the in-paralog and sister paralog analysis (see Chapters 3.2.4 and 3.2.5 respectively). The first test checks the percentage of OTUs remaining for each gene relative to the total number of OTUs selected by the user for inclusion in the analysis. After sequences identified by the distance matrix outlier analysis and in-paralog and sister paralog analysis have been removed, more than 75% of the OTUs must still be present to pass this test. Next the number of non-top hits are counted and an

average hits per OTU is calculated. A gene must have fewer than 1.1 hits per OTU to pass this test. Lastly the percentage of top hits flagged by the distance matrix outlier analysis and in-paralog and sister paralog analysis must be below 15%. It is possible for the user to set these cutoffs to values of even greater stringency. The highest stringency would be 100% of the OTUs present, 1.0 hits per OTU, and zero sequences should be flagged by the high stringency distance matrix analysis and in-paralog analysis. A new gene list is written and a command prompt is presented to the user to generate new alignments for tree finding using both the new gene list and excluding sequences flagged by the tests.

### 3.2.8 Preparation of Input Sequences for Case Study

Sequences for the case study came from two sources: (1) genomic assemblies and annotations deposited in The National Center for Biotechnology Information (NCBI), and (2) Illumina RNA reads from The Marine Microbial Eukaryote Transcriptome Sequences Project (MMETSP)[79] (Table 1).

Amino acid sequences and corresponding nucleic acid sequences were prepared for the genomic assemblies by extracting open-reading frames (ORFs) using the corresponding annotation file and the gffread utility provided in the Cufflinks package v2.2.1 and translating the ORFs using the TransDecoder v2.1.0 utility [80,81]. While many Stramenopile genome assemblies are available via NCBI, only those with annotations sufficient to extract ORFs were used in this case study.

Amino acid and corresponding nucleic acid sequences were prepared by assembling raw Illumina reads provided by the MMETSP. Prior to assembly, Illumina reads from the same species isolate were concatenated, then trimmed of sequencing adaptors and low quality sequences using the BBDuk package (09/2014) using a kmer size of 25 and trim quality cutoff of

10. The trimmed reads were then assembled using Trinity (version 2014 07 17), and ORFs were found using TransDecoder v2.1.0[79,81].

Each OTU was named according to standard binomial nomenclature, as provided by the submitter, plus an additional strain name when needed to differentiate between strains of the same species. Four taxa submitted to MMETSP contained either no identification beyond placement among Stramenopiles (one species), or placement only to the level of Order (three species). These taxa were given binomial names indicative of their lowest identified taxonomic level in order to have unique identification in the binomial structure expected by the pipeline.

### 3.3 Results and Discussion

#### 3.3.1 Case Study Input

An example analysis utilizing publicly available data was used to demonstrate the function and utility of the pipeline described here. One hundred sixty-four datasets comprised of 142 transcriptomic and 22 genomic data sets representing 127 species, 13 classes, and 43 orders across Stramenopiles (Supplemental Table 1) were analyzed to determine the phylogenetic affinity of seven taxa whose placement was uncertain (Table 1). It is worth noting that taxon sampling is poor amongst the *Phaeista*, Pseudofungi, and *Bigyra*.

Three of the unknown taxa, *Ochromonas*\_sp\_CCMP1899, *Ochromonas*\_sp\_CCMP1393, *Ochromonas*\_sp\_BG1, had been provisionally identified to the level of Genus, another three taxa, *Pelagophyceae*\_sp\_CCMP2097, *Pelagophyceae*\_sp\_RCC1024, *Pedinellales*\_sp\_CCMP2098, had predicted taxonomic placement to the level of Order, one taxon, *Stramenopile*\_sp\_NY07348D, had only been identified as a Stramenopile.

Library ID	Strain	Name	Phylum	Class	Order
MMETSP0198,9	LLF1b	Thraustochytrium_sp	Bigyra	Labyrinthulea	Thraustochytriida
MMETSP1105	BG1	Ochromonas_sp_BG1	Ochrophyta	Chrysophyceae	Ochromonadales
MMETSP0004,5	CCMP1393	Ochromonas_sp_CCMP1393	Ochrophyta	Chrysophyceae	Ochromonadales
MMETSP1177	CCMP1899	Ochromonas_sp_CCMP1899	Ochrophyta	Chrysophyceae	Ochromonadales
MMETSP0990-3	CCMP2098	Pedinellales_sp_CCMP2098	Ochrophyta	Dictyochophyceae	Pedinellales
MMETSP0974-7	CCMP2097	Pelagophyceae_sp_CCMP2097	Ochrophyta	Dictyochophyceae	Pelagomonadales
MMETSP1329	RCC1024	Pelagophyceae_sp_RCC1024	Ochrophyta	Dictyochophyceae	Pelagomonadales
MMETSP1433	NY07348D	Stramenopile_sp_NY07348D			

**Table 1 – Species with uncertain identification.**

Table of the seven taxa of uncertain phylogenetic placement. All unknown taxa were assembled from raw reads generated by the Marine Microbial Eukaryote Transcriptomic Sequencing Project (MMETSP).

### 3.3.2 Case Study Statistics

The pipeline was executed using the 164 transcriptomic and genomic data sets according to the default settings. The initial round of searches involved 495,116 uBLAST searches, yielding 342,535 unique hits. 88,261 of those hits passed the hmmsearch bitscore cut-off test, of which 68,642 were top hits leaving another 19,619 non-top hits. Upon completion of the first round of searches, the gene coverage table was manually reviewed and genes and species were successively culled until a final set of genes and species were selected in which 70% of the genes were found for each species and 70% of the species were present for each gene. This selection process yielded 150 OTUs of the original 164 analyzed and 387 genes of the original 603 to go forward in the pipeline.

New query terms for the second round of searches were produced from a set of 45,992 sequences, averaging 119 sequences per gene. Distance matrix outlier analysis identified 165 sequences for removal from the target databases. An additional 8,526 sequences were removed

from these databases based on the results of the in-paralog and sister paralog analysis. This amounted to 9.8% of the total hits after hmmsearch bitscore cut-off score screening. The second round of searches involved 6,898,800 uBLAST searches, yielding 378,821 unique hits, 60,846 of which passed the cut-off score test, for a total of 54,132 top hits and 6,714 non-top hits. Upon completion of the second round of searches the coverage table was again screened to produce a set of genes and species with very few data gaps. This produced a set of 150 species with at least 70% of genes present and 377 genes with at least 70% of the species present. The distance matrix analysis performed as part of the final tests revealed 85 putative paralogs, while the in-paralog analysis identified 920 putative paralogs. Following the final tests (see Chapter 3.2.7), a set of 373 genes were selected to find the final trees.

### 3.3.3 Case Study Phylogenetics

The Maximum-Likelihood (ML) trees generated by the long-branch outlier analysis provide an early look at how the pipeline is performing prior to any screening. The amino acid and nucleic acid trees are nearly identical (Supplemental Figure 1 and Supplemental Figure 2), differing largely in the branch order among *Cafeteria* sp., *Cafeteria roenbergensis*, and *Biscoid* sp., and the placement of a clade containing the *Cymatosirales* and *Triceratiales*. All unknown OTUs of interest have been placed in identical locations on both trees with bootstrap support greater than 80% in all cases. The three *Ochromonas* species were placed into a clade comprised of the genera *Ochromonas*, *Chromulina*, *Dinobryon*, *Paraphysomonas*, and *Mallomonas*, but not sister to one another. This clade contained another of the unknown species, *Pedinellales*\_sp\_CCMP2098. The two unknown *Pelagophyceae* species, *Pelagophyceae*\_sp\_CCMP2097 and *Pelagophyceae*\_sp\_RCC1024, were both found within the *Pelagomonadales*, but not sister to one another. *Pelagophyceae*\_sp\_CCMP2097 is placed basally

to the other *Pelagomonadales*, comprised of the genera *Aureococcus* and *Pelagomonas*. *Pelagophyceae*\_sp\_RCC1024 is sister to a clade comprised of the genus *Pelagomonas*. All bootstrap values within the *Pelagomonadales* are 100% for both the nucleic acid and amino acid trees. Lastly, the OTU identified as *Stramenopile*\_sp\_NY07348D was placed with 100% bootstrap support among the *Thraustochytriida* in both trees. One OTU, *Skeletonema*\_marinoi\_skelA was placed in an unexpected place with 90% bootstrap support in the amino acid tree and 100% bootstrap support in the nucleic acid tree. This OTU is found amongst the *Ochromonadales* rather than with the ten other members of its genus which form a distinct clade. It is worth noting that at this point in the pipeline non-orthologous sequences, including contaminant sequences, have not yet been screened.



Final phylogenetic analyses produce amino acid and nucleic acid trees nearly identical to the trees found after the first round of searches (Figure 11 and

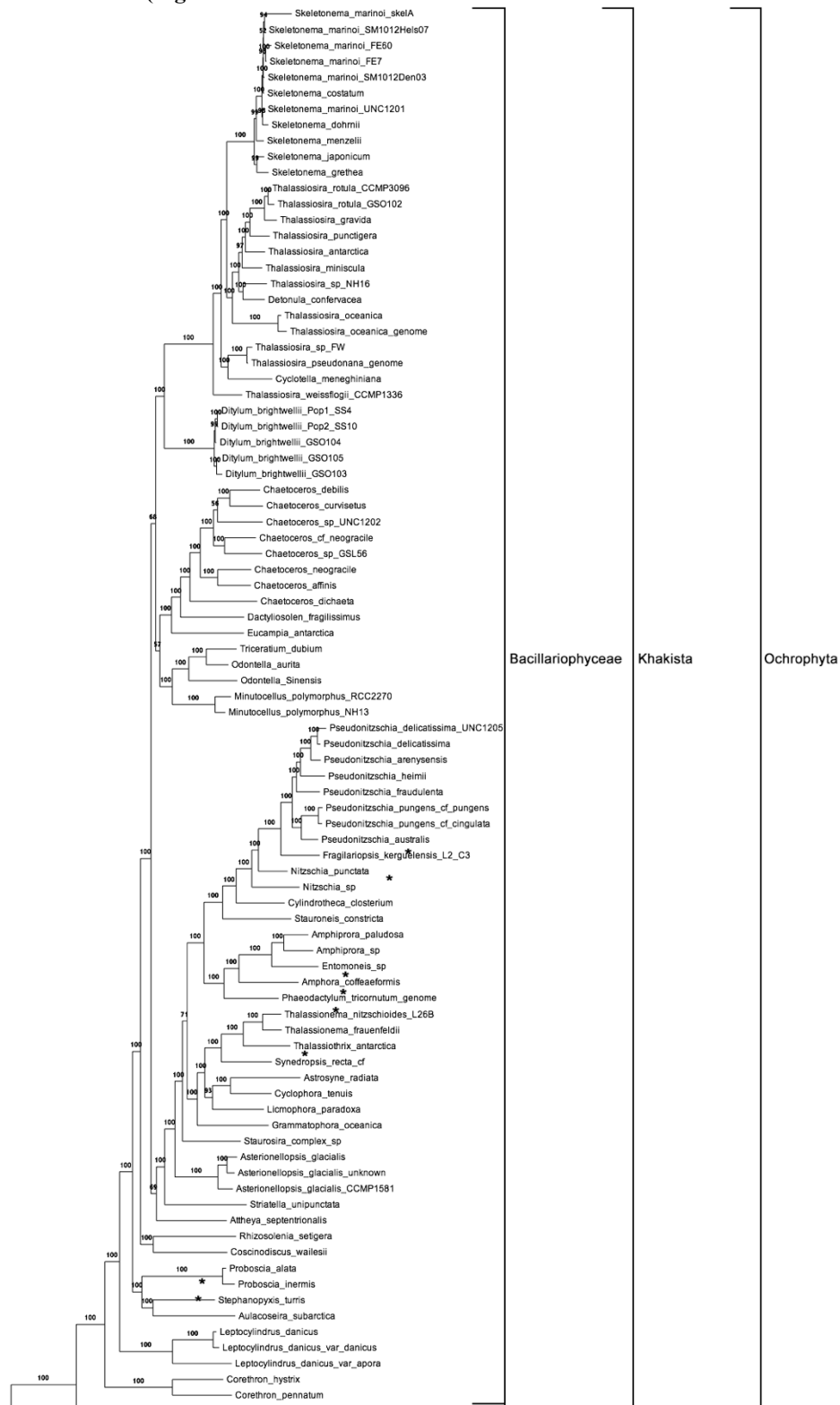
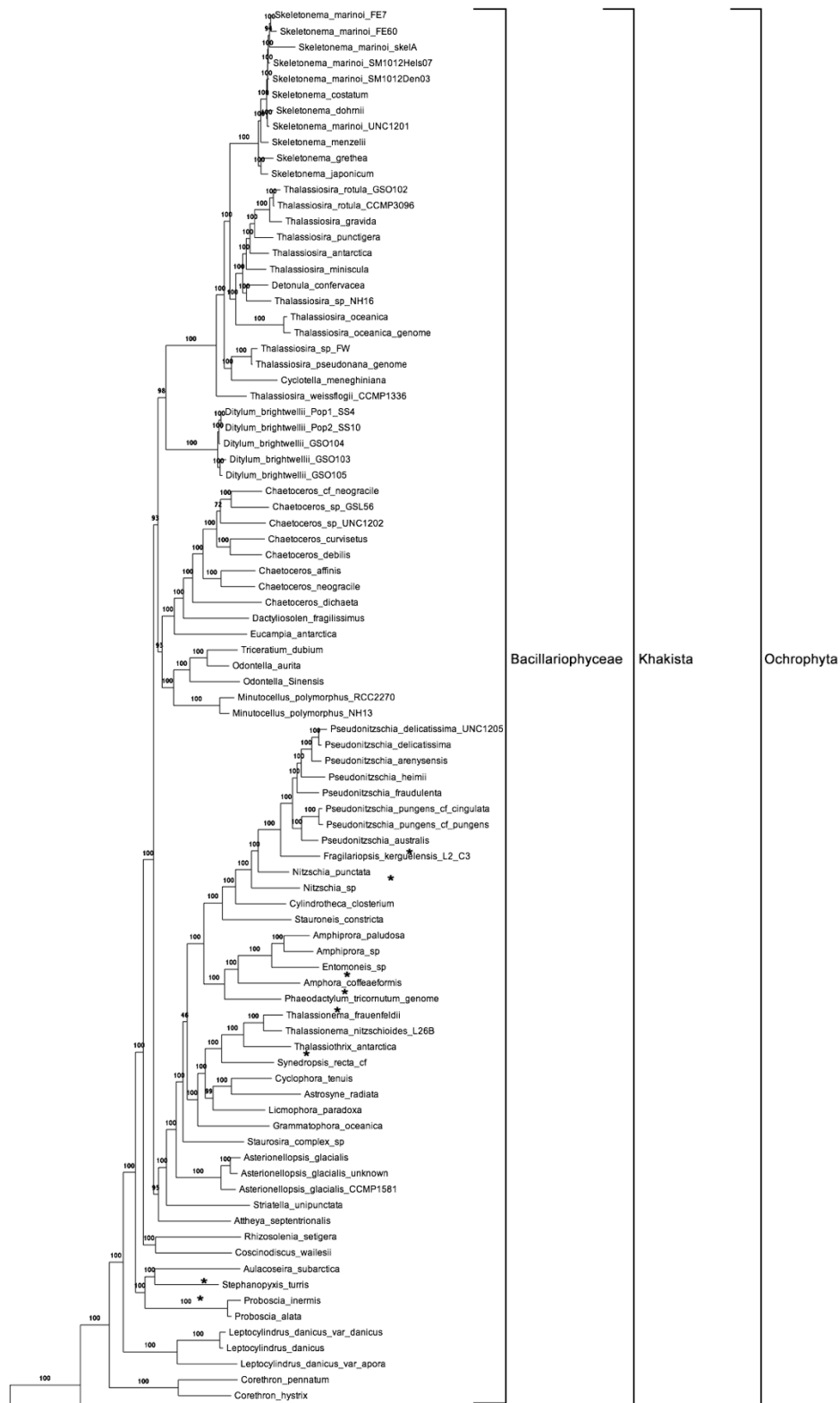


Figure 12). The *Cymatosirales* and *Triceratiales* form a monophyletic group in both trees, matching the topology of the pre- screening amino acid tree. The phylogenetic placement of the unknown species only changed in subtle ways. The biggest changes are amongst the *Ochromonas* spp. and *Pedinellales*\_sp\_CCMP2098. *Pedinellales*\_sp\_CCMP2098 is the earliest branch of a clade comprising the *Ochromonas*, *Dinobryon*, *Chromulina*, *Paraphysomonas*, and *Mallmonas*. The unknown species of *Ochromonas* form a clade with *Dinobryon* and *Chromulina*. The anomalous placement of *Skeletonema\_marinoi\_skelA* changed dramatically. In the final trees it is placed on a long branch among the other members of its genus. This change could be attributed to the removal of contaminant sequences during the screening. Bootstrap values across both trees worsened, most notably among the unknown *Ochromonas*, as well as in branches separating the non-monophyletic *Marista* and *Limnista* clades.

The *Phaeista* are paraphyletic, as are the *Marista*, and *Limnista* contained within it. The *Marista* are split amongst two clades, one comprising the *Alophyceae*, and another uniting the poorly sampled *Raphidophyceae*, *Phaeophyceae*, and *Xanthophyceae*. The placement of the *Pingulophyceae* differs between the nucleic acid and amino acid trees; placed at the base of the *Limnista* in the amino acid tree and at the base of the *Marista* in the nucleic acid tree. The *Limnista* are also split amongst two clades. One clade comprised of *Ochromonas*, *Chromulina*, *Dinobryon*, *Paraphysomonas*, and *Mallmonas*, and another comprised of *Pedinellales*, *Florenciellales*, and *Rhizosoleniales*. Poor bootstrap support at critical branches underlying these groups casts doubt on the accuracy of this topology, and it is notable that this is one of the areas with the poorest taxon sampling in this data-set.



**Figure 11 – Stramenopile Nucleic Acid Maximum Likelihood Tree**

The most likely tree found using RAXML using GTR substitution matrix with gamma correction and 100 bootstraps.

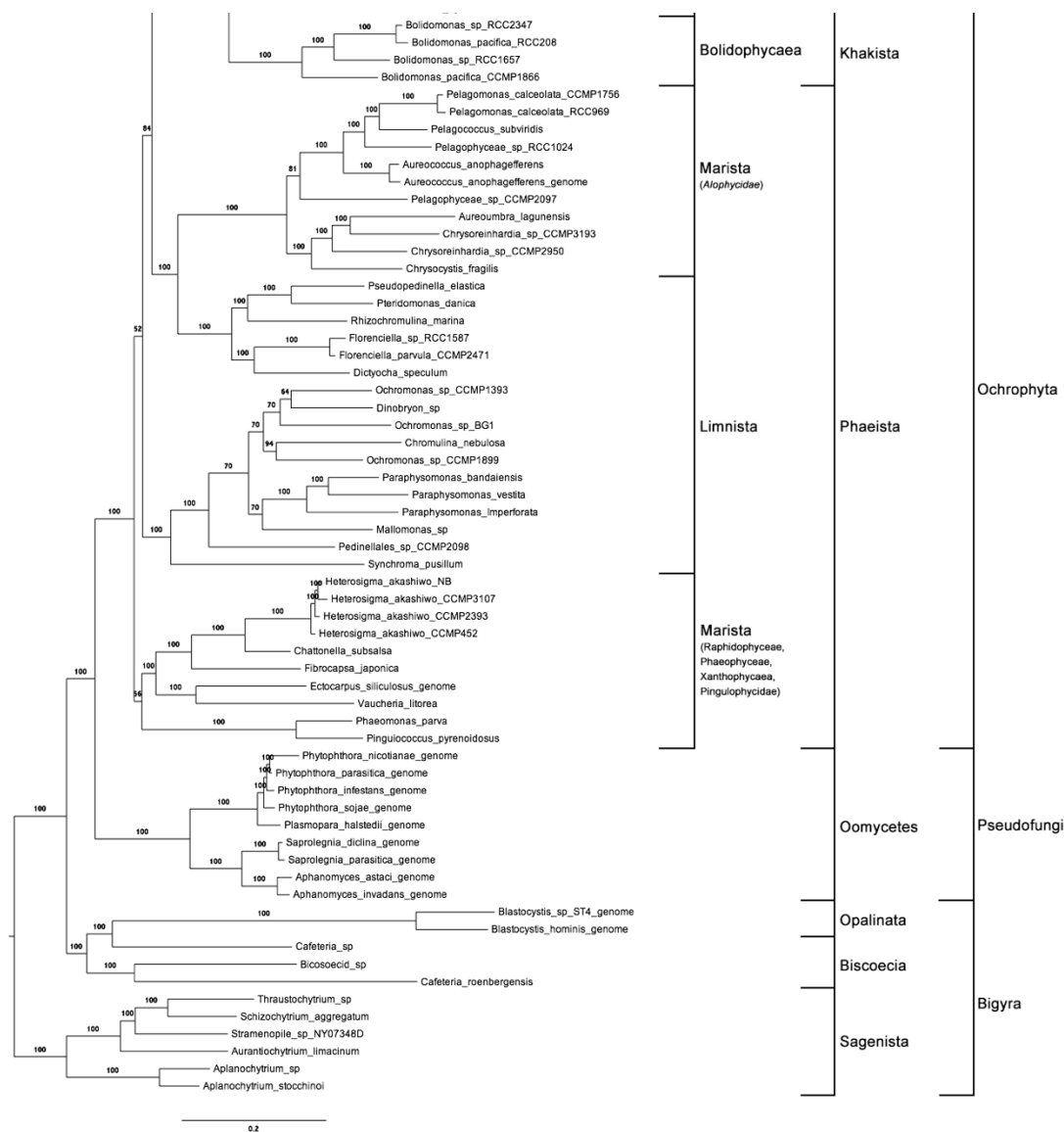
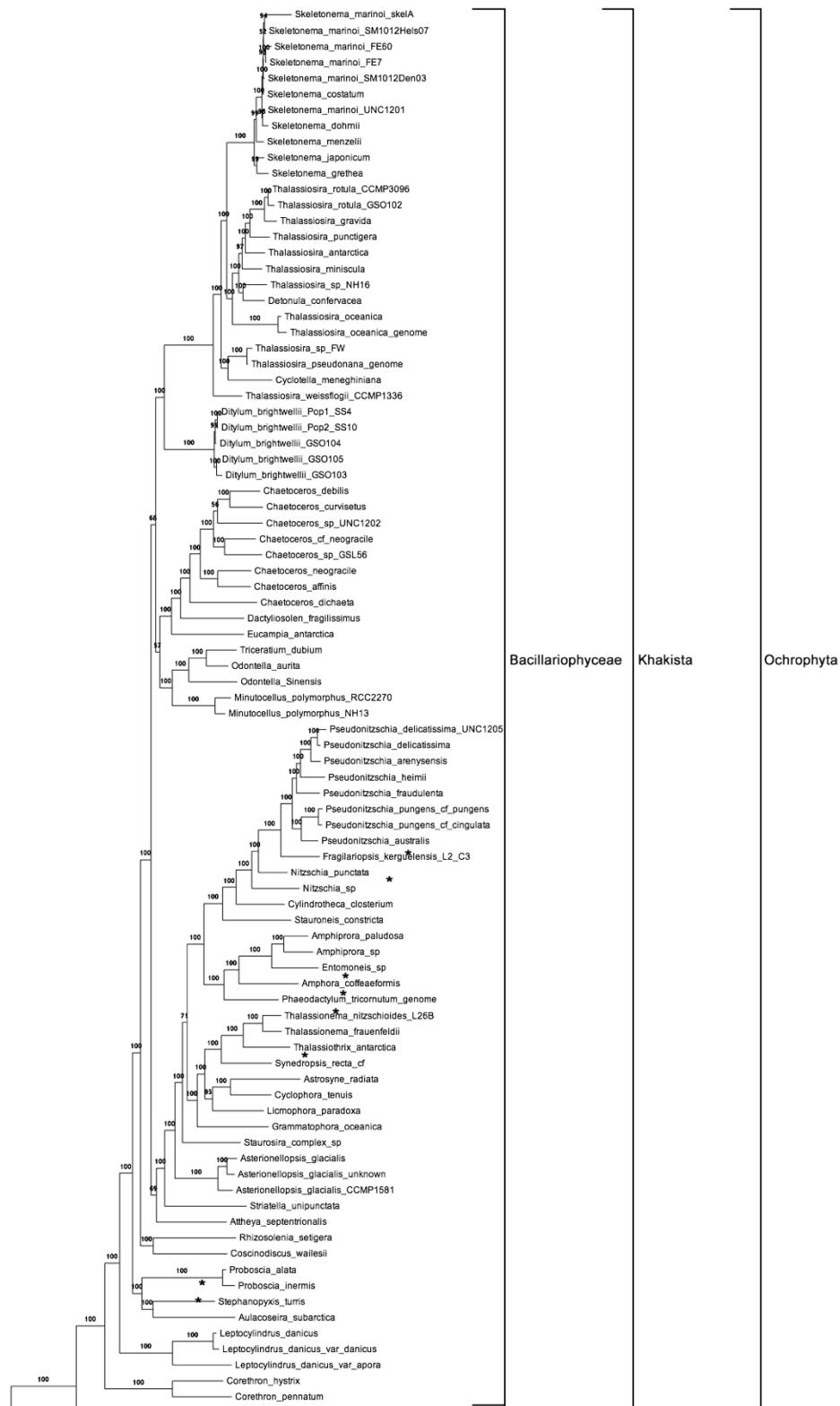


Figure 11 - Continued



**Figure 12 – Stramenopile Amino Acid Maximum Likelihood Tree**

The most likely tree found using RAxML using automatic selection of a substitution matrix with gamma correction and 100 bootstraps.

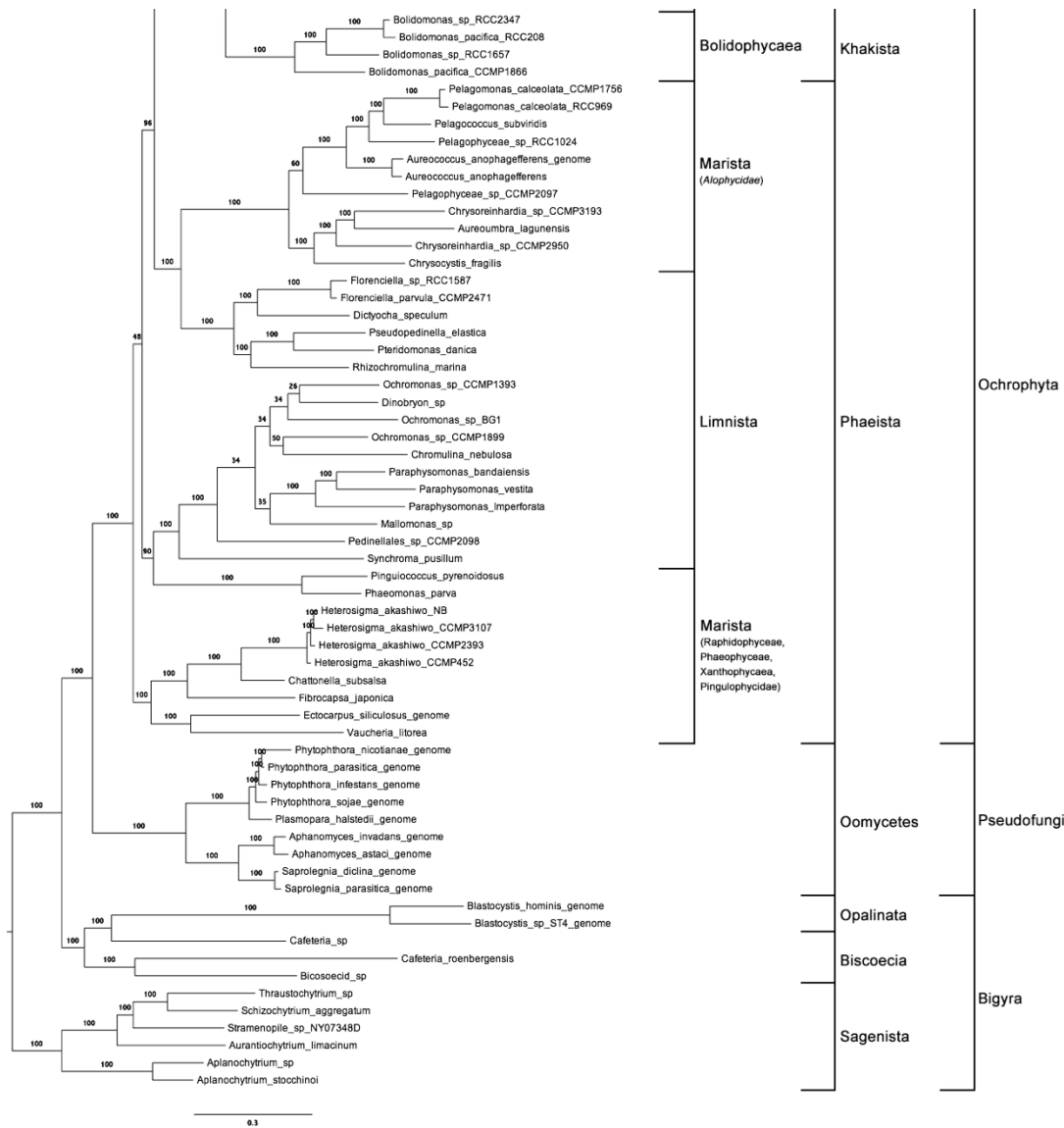


Figure 12 - Continued

### 3.4 Conclusions

The scripts and algorithms developed for this pipeline will be useful, either individually or in concert as part of the intended pipeline, to researchers interested in ortholog identification and phylogenetic analysis in cases where deep sequencing data is available for a large number of related species. The use of preliminary sequences from large number of deeply sequences taxa identified using traditional ortholog discovery methods to generate statistics to refine the identification of orthologs will be particularly useful in cases where complex gene families or contaminating sequences are expected to be present. The distance matrix outlier analysis in particular, may lend itself to more uses than applied by this pipeline, as it is a rapid taxonomy-agnostic method of identifying anomalous sequences in a set of amino acid or nucleic acid sequences.

This pipeline greatly improves the identification of orthologs in groups with deep divergence times or complex gene families that may hamper traditional ortholog clustering techniques. While many phylogenetic pipelines have been made available that utilize deep sequencing data [69,70,82], my pipeline is novel in that it leverages statistical analyses of all clusters under scrutiny to provide an additional layer of information beyond pairwise comparisons of single target and query sequences or target to sequence models. Because the pipeline is so dependent on deep sequencing of multiple related species, it fails to provide benefits beyond traditional ortholog clustering when taxon sampling is sparse. This failing is particularly noticeable among the Bigyra, where sparse taxon sampling has resulted in long branches uniting OTUs. Without broad taxon sampling, the pipeline is only a modest improvement over traditional forms of ortholog detection. Further, since the distance matrix analysis makes use of data from all sequences identified by the initial ortholog search, the

analysis could be greatly improved with the addition of more gene queries. Since the CEGMA and BUSCO query terms were generated as conserved clusters across all eukaryotes, there are likely additional clusters that may be specific to the lineages under consideration. The addition of lineage specific clusters would increase the depth of data under analysis and thereby improve the statistical analysis underlying the distance matrix analysis and thus improve the overall functioning of the pipeline.

### 3.5 *Availability of supporting data*

All transcriptomic data can be found at the iMicrobe website under MMETSP (<http://data.imicrobe.us/project/view/104>). Genomic assemblies were downloaded from Genbank or JGI as indicated in Supplemental Table 1. Scripts developed for this pipeline, and detailed documentation can be found at the following github repository: <https://github.com/mendezg/DATOL>.



## 4 Application of a New Ortholog Detection Pipeline to the Discovery of Phylogenetic Markers in Dinoflagellates

### 4.1 Introduction

Dinoflagellates are an important group of microorganisms whose members take on trophic roles as diverse as primary producers, predators, and parasites, and even directly affect human health in the form of harmful algal blooms. They are a large diverse group of organisms, estimated to number 11,000 species[83]. In marine systems they make up the most abundant group below 5 $\mu$ m, and the most species rich group below 180 $\mu$ m[84]. Taylor's 1987 synthesis of phylogenetic and morphological data, identifies six major groups: *Gymnodinales*, *Peridinales*, *Suessiales*, *Gonyaulacales*, *Prorocentrales*, and *Dinophysiales*[23]. These initial six clades of dinoflagellates described by Taylor[23] all unambiguously exhibit a dinokaryon, permanently condensed chromosomes lacking nucleosomes, and make up the dinophyceae. Later analyses by Taylor [24] and Fensome[25], added the *Syndinales* and *Noctilucales*, which were thought to lack a dinokaryon all the time or only part of the time (respectively). Study of the Syndinales species *Amoebophrya* would later revise this understanding when a dinokaryon was observed during specific times during its life cycle[26]. While the Syndinales and Noctilucales are considered members of *dinokaryota*, their cryptic expression of the dinokaryon has placed them outside of the dinophyceae in syntheses of phylogenetic and morphological data[24,27]. Unfortunately, the Noctilucales and Syndinales (as well as the Dinophysiales) are difficult to culture, and therefore molecular data of individuals from these groups are difficult to obtain. Studies of this diverse group of organisms has struggled to elucidate the relationships between the major taxonomic groups, as well as the placement of many individual species. Traditional

morphological analyses have relied on the arrangement of cellulosic plates located in cortical alveoli, termed thecal plates, but many dinoflagellates lack thecal plates and are thus assigned to a clade, the Gymnodinales, comprised of species whose defining characteristic is a lack of this common morphological character (though a laborious procedure involving the stripping of the outer membrane can be used to investigate the arrangement of the cortical alveolae). Further vexing morphological taxonomy are clades where there has been significant modification of thecal plates, such as the Prorocentrales, where two plates, termed valves in this clade, cover the majority of the cell surface and the remaining plates have been reduced to microplates around an apical pore from which the organism's two flagella emerge. Phylogenetic analyses have dramatically improved the taxonomic understanding of dinoflagellates, most notably by moving them from a dubious position outside *Eukaryota* to a position in the crown of Eukaryota sister to the parasitic *Apicomplexa*[2-4]. However, a series of unusual genomic characteristics has presented significant challenges to phylogeneticists. The most significant of these unusual genomic characteristics are:

1. ribosomal DNA (rDNA) is a poor phylogenetic marker, resulting in poorly supported clades that change greatly depending on the taxa selected for the analysis.
2. The mitochondrial and chloroplast genomes are small and unusual in structure
3. The nuclear genome size averages 20X the human genome.
4. Most, if not all, genes occur in large complex gene families

Difficulties in culturing dinoflagellates and in amplifying genes of interest from dinoflagellates has meant that phylogenetic studies utilizing more than one gene remain rare, and taxon sampling continues to hamper analyses. While taxon sampling is improving, most analyses still utilize only rDNA genes (usually the small and large subunit of the nuclear operon; SSU or

LSU). Of the eight clades noted here, only two of them, the Suessiales and Dinophysiales, reproducibly form monophyletic groups in phylogenetic analyses of rDNA, though it is rare for multiple members of any of these groups to be present in rDNA analyses[5-9]. The addition of mitochondrial or chloroplast genes to phylogenetic analyses has been hindered by massive transfer of genes from these organellar genomes to the nuclear genome and rapid sequence evolution in these genomes[10-12]. An unresolved knot of species classified in different groups based upon morphological characters, termed the GPP complex and comprised of members of the Gymnodinales, Peridinales, and Prorocentrales, was partially unraveled with the addition of protein coding genes to phylogenetic analyses [25,28,39,85-87]. Taxon sampling in these larger phylogenetic analyses was likely limited by the difficulties isolating orthologous sequences from a large number of dinoflagellate taxa, owing to large genomes and large complex gene families of dinoflagellates. Genomic sequences are unusually difficult to obtain owing to genomes that are on average 10X larger than the human genome[13], and with most genes belonging to large heterogeneous gene families[19]. Collectively, these unusual characteristics significantly impede phylogenetic studies. One recent phylogenetic analysis utilized a set of 73 short protein-coding ribosomal genes identified from a mixture of 19 dinoflagellate expressed sequence tag (EST) libraries and new transcriptomic assemblies[28]. That analysis consistently supported monophyly for the Gonyaulacales, Prorocentrales, and Suessiales present, but was unable to resolve relationships between major clades or place *Cryptothecodinium cohnii*, a species with uncertain phylogenetic affinity. Despite utilizing all deep sequencing data available at the time, the study was likely hindered by taxon sampling, notably lacking any members of the Peridinales, Dinophysiales, or Noctilucales.

Here a custom pipeline is used to identify orthologs given the unusual challenges of the dinoflagellate nucleus. 668 genes are identified comprising nearly 700,000 base pairs. Representatives of all eight major dinoflagellate clades are present, and multiple representatives from five of the major clades, Gymnodinales, Suessiales, Prorocentrales, Gonyaulacales, and Peridinales, are present. This allows for phylogenetic analysis on a scope not previously possible for this group of organisms, and sheds light on the evolution and relationships within this large and diverse group.

## 4.2 Materials and Methods

### 4.2.1 Taxon Selection and Transcriptomic Assembly

Seventy-seven transcriptomic assemblies and one genomic assembly were selected for analysis, representing 54 dinoflagellate species across seven major clades, and three outgroup taxa (Supplemental Table 1). To our knowledge, this was all transcriptomic and genomic scale data available at the time this work was performed. The data from *Perkinsus marina* is the only genomic dataset in this analysis. Raw reads from 69 of the assemblies was acquired from the Marine Microbe Eukaryotic Transcriptomic Sequencing Project (MMETSP). Eight transcriptomes were sequenced in-house using the culturing, RNA extraction, and sequencing methods as previously described [28]. The transcriptomic assembly from *Hematodinium* sp. was provided by Ross Waller [88]. These data cover 54 dinoflagellate species, with 11 species represented multiple times by separate isolates. The three non-dinoflagellate species, *Perkinsus marina*, *Vitrella brassicaformis* (two isolates), and *Chromera velia*, were selected for use as outgroups since they are thought to belong to groups between dinoflagellates and their sister clade the Apicomplexans.

Raw Illumina reads from MMETSP and those produced in-house were first pooled, where possible, so that separate sequencing runs from identical isolates would be concatenated into one assembly. These reads files were then processed using BBduk(09/2014) to remove Illumina adaptor sequences and trim low quality sequences using a kmer size of 25 and trim quality cutoff of 10. Processed reads were assembled using Trinity (version 2014 07 17). Open-reading frames from transcriptomic assemblies were identified and written using TransDecoder v2.1.0.

#### 4.2.2 Query Sequence Preparation

Query terms for ortholog identification were combined from three non-redundant sets of highly conserved gene clusters. As previously described (see Chapter 3), query terms, HMMs, and HMMsearch bitscore cut-offs from the CEGMA and BUSCO projects were combined into a non-redundant set of 729 query terms. For this analysis, a third set of query terms was added to the CEGMA and BUSCO set. Amino acid sequence files from eight dinoflagellate transcriptomic assemblies were selected representing five major dinoflagellate clades (Table 1).

OTU	Clade	Sequences
<i>Amphidinium_carterae_MMET</i>	<i>Gymnodinales</i>	44,771
<i>Amphidinium_massartii</i>	<i>Gymnodinales</i>	54,308
<i>Gambierdiscus_australes</i>	<i>Gonyaucales</i>	67,226
<i>Gonyaulax_spinifera</i>	<i>Gonyaucales</i>	45,056
<i>Gyrodinium_instriatum</i>	<i>Gymnodinales</i>	181,267
<i>Peridinium_aciculiferum</i>	<i>Peridinales</i>	81,349
<i>Prorocentrum_minimum</i>	<i>Prorocentrales</i>	109,764
<i>Symbiodinium_sp_CCMP421</i>	<i>Suessiales</i>	93,175

**Table 2 – Assemblies used for Clustering**

Transcriptomic assemblies selected for preliminary ortholog clustering to generate query terms to be used in the more extensive ortholog discovery pipeline.

Sequences were clustered using get\_homologues [89] using the COGtriangles [90] algorithm and configured to require at least one sequence from each species. The resultant 1486 clusters were aligned using MAFFT (version 7.215) [73], and Maximum Likelihood trees were found using RAxML v8.1.24 using automatic protein model selection and gamma correction[75]. Trees were analyzed using custom python scripts to test whether cases of multiple sequences from the same OTU yielded a monophyletic group. Clusters where all sequences from each species were monophyletic were selected as potential query terms for use in the pipeline. Each of these dinoflagellate clusters was then checked for similarity to query terms already in the non-redundant set of 729 CEGMA and BUSCO query terms. Only sequences with no hits to sequences in the CEGMA and BUSCO sets were retained as query terms for the pipeline. This process yielded 763 dinoflagellate sequence clusters. When combined with the 729 query terms

from CEGMA and BUSCO this is a total of 1492 query clusters. HMMs were built for each dinoflagellate cluster using HMMbuild 3.1b1[63] and bitscore cut-offs were calculated as previously described (see Chapter 3.2.2).

#### 4.2.3 Ortholog Identification

Putative orthologs were identified using the custom pipeline previously described (see Chapter 3) using default settings for all variables, and query terms described above (see Chapter 4.2.2). After the searches of each stage of the pipeline, the resultant coverage table was reviewed as instructed by the pipeline to generate a list of OTUs and genes to move forward in the pipeline. To generate a less “gappy” dataset, species and genes with poor coverage were successively culled, with live updating of the remaining percentages, until all species had at least 70% of the genes and all genes were present in at least 70% of the species.

#### 4.2.4 Phylogenetic Analysis

Analyses were run using RAxML version 8.1.17 [75]. Nucleic acid analyses were performed using a general time reversible model with gamma rate correction. Amino acid analyses were performed using automatic protein substitution model selection with gamma rate correction. For branch support of concatenated multiple sequence alignment analysis trees, nonparametric bootstrap analyses were performed with 100 replicates.

### 4.3 Results

#### 4.3.1 Sequence Searches and Screening

The initial round of searches involved 942,474 uBLAST searches, yielding 559,706 unique hits. 111,197 of those hits passed the hmmsearch bitscore cut-off test, of which 68,614 are top hits leaving another 42,583 non-top hits. The manual review of the coverage table

produced a set of 58 species and 1,125 genes to continue forward in the pipeline. New query terms for a second round of searches were produced from a set of 50,577 sequences, averaging 45 sequences per gene. Distance matrix outlier analysis identified 839 sequences for removal from target databases. An additional 27,522 sequences were removed from these databases based on the results of the in-paralog and sister paralog analysis. This amounts to 25% of the total hits after hmmsearch bitscore cut-off score screening. The second round of searches involved 3,945,006 uBLAST searches, yielding 575,828 unique hits. This was reduced to 75,005 following the cut-off score screening, for a total of 60,846 top hits and 14,159 non-top hits. The coverage table review produced a set of 58 species and 1,101 genes for final testing. The distance matrix outlier analysis detected 87 putatively non-orthologous sequences, and the in-paralog and sister paralog analysis identified 3,484 putatively non-orthologous sequences. Of the 1,101 genes examined in the final analysis, a set of 668 genes were recommended for phylogenetic analysis.

#### 4.3.2 Phylogenetic Analyses

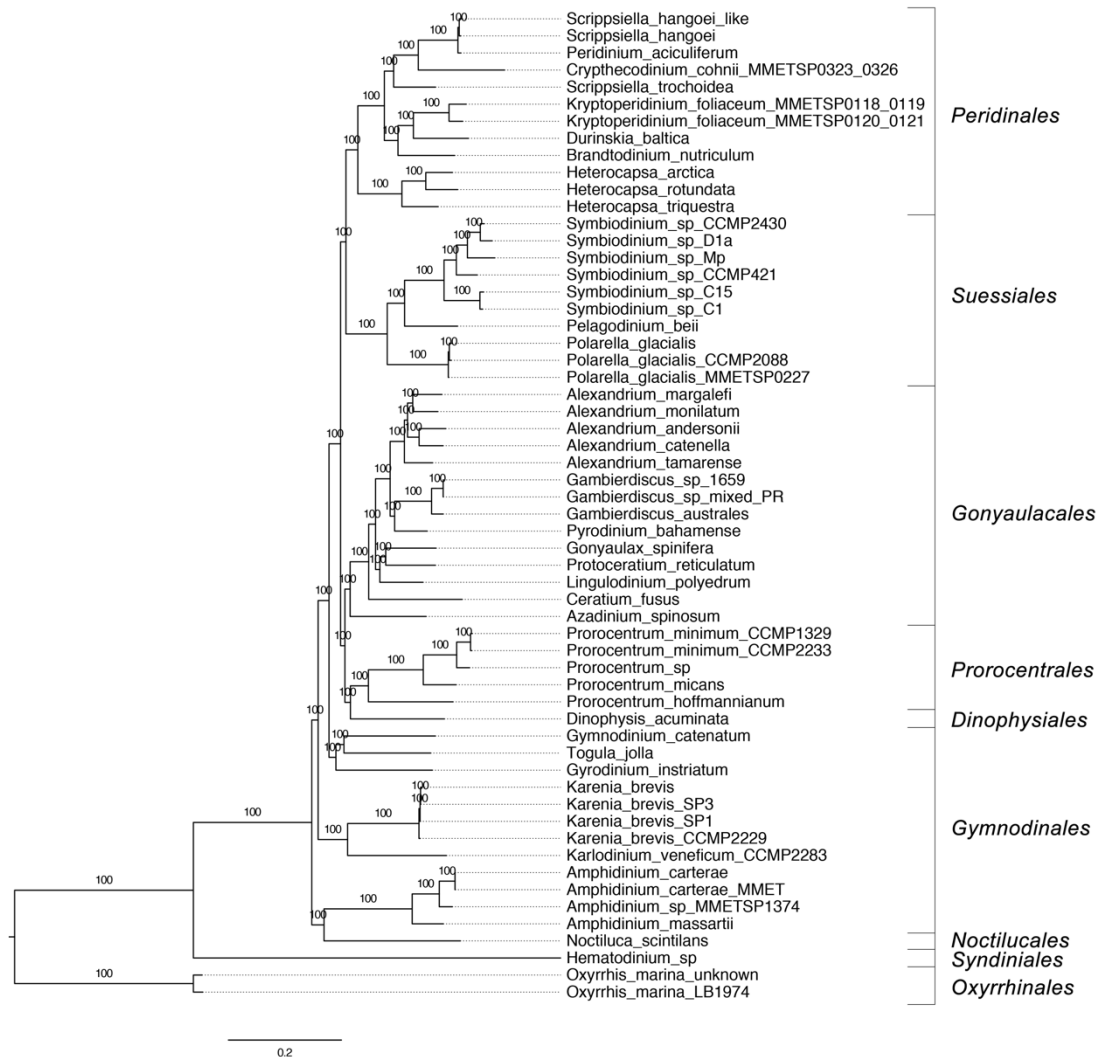
The concatenated MSA amino acid and nucleic acid trees produced prior to screening show little disagreement between the two trees (Supplemental Figures 1 and 2). The placement of *Noctiluca scintilans* is the only topological difference between the two trees; sister to the Amphidinales in the nucleic acid tree and moved one tree node more basal in the amino acid tree to be sister to all dinophycean dinoflagellates. Of the major taxonomic clades as discussed by Taylor for which multiple species are present in the analysis, only the Gymnodinales fails to form a monophyletic group[24]. The Suessiales, Gonyaulacales, Prorocentrales, and Peridinales each form monophyletic groups. Taxa of unknown phylogenetic affinity, *Crypthecodinium cohnii*, *Heterocapsa* spp, *Azadinium spinosum*, and *Dinophysis acuminata*, have been placed on the tree with 100% bootstrap support in all cases other than a 73% bootstrap support on the



branch of *Azadinium spinosum* on the nucleic acid tree. Two branches on the amino acid tree are below 100% bootstrap support; the branch placing *Alexandrium monilatum* and *Alexandrium margalefi* sister to *Alexandrium tamarense* (95%), and the branch placing *Togula jolla* and *Gymnodinium catenatum* sister to *Gyrodinium instriatum* (93%) (Supplemental Figure 4). In both cases, short internal branches link long terminal branches. Four branches have bootstraps below 100% on the nucleic acid tree: the branch placing *Azadinium spinosum* sister to the Suessiales (73%), The branch placing the *Azadinium spinosum* – Suessiales clade sister to the Peridinales clade (73%), the branch placing the Prorocentrales – Dinophyiales clade sister to the Gonyaulacales (73%), and the branch placing *Gymnodinium catenatum* sister to *Togula jolla* (99%) (Supplemental Figure 3).

The final concatenated multiple sequence alignments included 58 OTUs and 668 genes. The nucleic acid alignment contained 699,010 characters of which 326,149 were parsimony-informative. The amino acid alignment contained 220,547 characters of which 143,963 were parsimony-informative. The topology of the final amino acid and nucleic acid trees is identical (Figure 13 and Figure 14). *Noctiluca scintillans* is sister to the Amphidinales, as it is in the nucleic acid tree generated prior to paralog screening. Bootstrap values for the nucleic acid tree are all 100%, while a single branch on the amino acid tree is below 100%; the branch placing *Noctiluca scintillans* sister to the *Amphidinales* has a bootstrap score of 98%. All major dinoflagellate clades are monophyletic, with the exception of the Gymnodinales. It is worth noting, however, that the Dinophysiales, Noctilucales and Syndinales are each represented by a single taxon, and hence cannot be tested for monophyly. The positions of *Crypthecodinium cohnii*, and the *Heterocapsa* species remain unchanged from the pre-screening trees, both members of the Peridinales. *Azadinium spinosum*, on the other hand, is placed as the earliest

branch among the Gonyaulacales. The Prorocentrales and Dinophysiales form a monophyletic group sister to the Gonyaulacales. The Suessiales are sister to the Peridinales, and this clade is sister to the clade comprised of the Gonyaulacales, Prorocentrales, and Dinophysiales. The Gymnodinales form three separate monophyletic groups: the Amphidinales (sister to the Noctilucales), the fucoxanthin-containing *Karenia* (*Karlodinium veneficum* and *Karenia Brevis*), and the remaining Gymnodinales (*Togula jolla*, *Gymnodinium catenatum*, and *Gyrodinium instriatum*). *Kryptoperidinium foliaceum* and *Durinskia baltica*, the dinoflagellates which contain a tertiary plastid of diatom origin form, a monophyletic group sister to *Brandtodinium nutriculum* within the Peridinales. The genus *Scropsiella* does not form a monophyletic group, with both *Peridinium aciculiferum* and *Crypthecodinium cohnii* within the smallest clade comprising all *Scropsiella* species.



**Figure 13 – Dinoflagellate Nucleic Acid Maximum Likelihood Tree**

The most likely tree found using RAxML using GTR substitution matrix with gamma correction and 100 bootstraps. Separate multiple sequence alignments (MSAs) of 668 genes were concatenated into a single MSA of 699,010 characters of which 326,149 were parsimony-informative



#### 4.4 Discussion

That dinoflagellate phylogenetics could be improved through increased taxon sampling and the addition of more protein coding genes to analyses has been apparent for some time and the focus of several previous studies [25,28,38,85-87,91]. While recent advances in sequencing have made transcriptome-scale data accessible for the first time, identifying orthologs in the unique genomic environment of the dinoflagellate cell is challenging. In this study, I introduce a novel pipeline that combines traditional ortholog discovery approaches with new algorithms that leverage the broad taxon sampling and deep sequencing data now available. As a result, the number of genes available to phylogenetic analysis has been greatly expanded. Taxonomic groups based primarily upon plate tabulation (Gonyaulacales, Peridinales, and Suessiales) have held together well in phylogenetic analyses including multiple genes, but here, the relationships between these groups are also well supported using sequencing data. While well supported relationships between major dinoflagellate lineages might be new, the underlying topology offers few surprises. In fact, the topology of the trees found in this study is perfectly consistent with the trees presented in Bachvaroff et al. 2014 [28].

The placement of *Cryptocodinium cohnii* and the genus *Heterocapsa* within the Peridinales, with strong bootstrap support, will hopefully resolve confusion surrounding the inconsistent placement of these groups, particularly as it relates to the treatment of these organisms as early branching dinoflagellates offering insights into the ancestral state of dinoflagellate characters. Instead, the paraphyletic group of non-dinophycean taxa are the earliest branching dinoflagellates. Dinophycean dinoflagellate, characterized by a permanent dinokaryon, are monophyletic. Improved taxon sampling among the difficult to culture Gymnodinales, Noctilucales, and Syndinales will be necessary to further explore questions

surrounding the ancestral states of dinoflagellate characters, but the topology here and of Bachvaroff [28] suggests a transition from an ancestral state in which a dinokaryon was only present during part of the life cycle to a permanent dinokaryon.

The evolution of cellulosic thecal plates has followed a similar evolutionary process. A paraphyletic group of athecate dinoflagellates occupy the base of the dinoflagellate tree, while the thecate dinoflagellates form a monophyletic clade. Within this thecate clade, both the reduction of thecal plates in groups such as the Dinophysiales and Prorocentrales, as well as the increase in thecal plates, as in the Suessiales, can be recognized as derived character states rather than indications of an ancestral state.

It has long been suspected that Dinophysiales and Prorocentrales are sister clades on the basis of plate tabulation [92], but until this study it has never been supported in a phylogenetic analyses. The presence of toxin producing species in both of these groups is interesting, and a deeper phylogenetic analysis of this group undertaken in the context of the presence of toxin producing genes across the clade could be helpful in identification and management of harmful algal blooms produced by members of this clade. The placement of this Prorocentrales-Dinophysiales group sister to the Gonyaulacales supports hypotheses of thecal plate reduction and anterior flagellation as derived characters in this group, which was previously suggested by Fensome [92].

Logares had previously noted the genetic similarities between *Scrippsiella hangoei* and *Peridinium aciculiferum*, suggesting the species diverged recently[93]. The non-monophyly of the genus *Scrippsiella* in my analysis as well as the similarity of sequences, here and noted previously by Logares[93], between *Scrippsiella hangoei* and *Peridinium aciculiferum* indicates the need for a deeper analysis and reevaluation of taxonomy of this group.

Other areas of this analysis could be greatly improved by increased taxon sampling. The Dinophysiales, Noctilucales, and Syndinales, each with only a single member representing the entire clade, are obvious areas that could use more representatives, but the bases of both the Gonyaulacales and Prorocentrales are also poorly sampled. In fact, only a single member of the benthic toxin producing clade of Prorocentrales, *Prorocentrum hoffmannianum*, is present. The classification of taxa placed in the Gymnodinales, revealed as paraphyletic, is another group ready for deeper analysis and reevaluation. This analysis suggests the Gymnodinales can be divided into three clades: one closely related to the thecate dinoflagellates, the Kareniaceae, and the Amphidinales. Although this is the most taxon-rich analysis of this size we are familiar with, it still has relatively few members of the *Gymnodinales*, and these three clades are likely to be revised in further studies.

#### 4.5 Availability of supporting data

All transcriptomic data can be found at the iMicrobe website under MMETSP (<http://data.imicrobe.us/project/view/104>). Genomic assemblies were downloaded from Genbank or JGI as indicated in Supplemental Table 2. Scripts used in this analysis, and detailed documentation can be found at the following github repository: <https://github.com/mendezg/DATOL>.

## 5 Conclusions

The unique characteristics of the dinoflagellate nucleus has long presented challenges for the evolutionary understanding of this group of organisms. The permanently condensed and fibrillary arched bands of dinoflagellate chromosomes led to mistaken interpretations of dinoflagellates as early branching eukaryotes or even a fourth kingdom of life entirely (the Mesokaryota)[1]. As phylogenetic techniques became available, the large genomes and tandemly repeated genes of the dinoflagellates would stymie phylogenetic studies [13,17,19]. Taxonomic groups based largely upon morphological analyses of thecal plate arrangement were supported by phylogenetic analyses of rDNA, but relationships between these taxonomic groups remained elusive[25,92,94]. An unresolved knot of species classified in different groups based upon morphological characters, termed the GPP complex and comprised of members of the Gymnodinales, Peridinales, and Prorocentrales, was partially unraveled with the addition of protein coding genes to phylogenetic analyses [25,28,39,85-87]. Taxon sampling in these larger phylogenetic analyses was likely limited by the difficulties isolating orthologous sequences from a large number of dinoflagellate taxa, owing to large genomes and large complex gene families of dinoflagellates. Recent advancements in sequencing technology has made deep transcriptomic sequencing of dinoflagellates widely accessible, but current methods of ortholog identification were not designed with the challenges of the dinoflagellate nucleus in mind, nor were they designed to leverage the statistical power these large datasets offer.



In this dissertation the longest contiguously sequenced tandem gene array is sequenced and analyzed. Current models of dinoflagellate genomic organization are described and compared to the tandem array. While this work was not designed to comprehensively test the or accuracy of existing models of genomic organization, it is notable that the genomic organization observed in this tandem repeat is inconsistent with existing models. This is particularly evident in the intron density of the tandem repeat. Tandem repeats had previously been described as intron-poor [19], but this tandem repeat had more intronic bases than exonic. Intron dense genes were expected to be present only in low copy genes, but this gene was the most numerous sequence in EST libraries of the species under study. The conservation of intergenic and intronic regions of the tandem repeat also suggest a genome-level duplication method as well as relatively recent duplication or concerted evolution (or both). The rapid advancements in sequencing technology and the associated reduction in the price per sequenced base will soon make the complete sequencing of a dinoflagellate genome economically feasible. Multiple attempts have already been made to survey the genomes of the dinoflagellates with the smallest genomes, the endosymbiotic *Symbiodinium* species [20,21]. Perhaps the most useful finding of the analysis of the tandem repeat was the characterization of a non-canonical splicing method. The description of this previously undescribed non-canonical splicing method will certainly improve modeling and automated annotation of genes in an assembled dinoflagellate genome. While genomic data for dinoflagellates is currently scarce, future phylogenetic studies will likely prefer to use whole genome data to

transcriptomic data. Accurate automated modeling and annotation of genes will be fundamental in the identification of orthologs from data of this kind.

To leverage the tremendous potential of new transcriptomic data sets, a new approach to ortholog discovery was developed. Building upon existing methods of ortholog detection, this pipeline employed statistics of the data-set encompassing many species and genes to screen the orthologs discovered using the standard techniques for anomalous sequences (contaminants, horizontally transferred genes, or paralogous sequences). In this manner, a much larger set of phylogenetic characters should be analyzed than what would be feasible with more manual approaches.

Phylogenetic analysis of dinoflagellates with the genes identified with this pipeline was consistent with the topology of the most recent phylogenetic studies of the group[28], but for the first time resolved relationships between major dinoflagellate clades with good bootstrap support and consistency between amino acid and nucleic acid trees. Organisms that had long held inconsistent or poorly supported positions on the tree, such as *Cryptothecodinium cohnii*, had well supported placement on the trees. The hypothesis placing the Dinophysiales sister to the Prorocentrales, long supported based on morphological characters [24,25,92] had strong support. Three paraphyletic clades of Gymnodinales were revealed, which underscores the need for improved taxon sampling of this under-sampled group of dinoflagellates. The strong support for a Suessiales-Peridinales clade and a Gonyaulacales-Prorocentrales-Dinophysiales clade, as well as a clade of Gymnodinales sister to this thecate clade is a hypothesis in need of further testing as well as detailed study in the context of the morphological characters.

The pipeline developed and employed in this dissertation was a necessary improvement over existing ortholog discovery tools, due to the unique challenges of the dinoflagellate nucleus. Notably, the initial round of sequence searches from the dinoflagellates yielded over 40% non-top hits, while the equivalent search in the Stramenopiles yielded 22% non-top hits. However, the pipeline's use of statistics generated from full genomes and transcriptomes from many species also proved useful in the discovery of orthologs in the Stramenopiles, a group of organisms with typical eukaryotic genome organization, size, and gene duplication. It is likely that approaches similar to those used in this dissertation may be helpful in the reconstruction of other groups of eukaryotes, and perhaps in resolving the relationships between major eukaryotic clades.

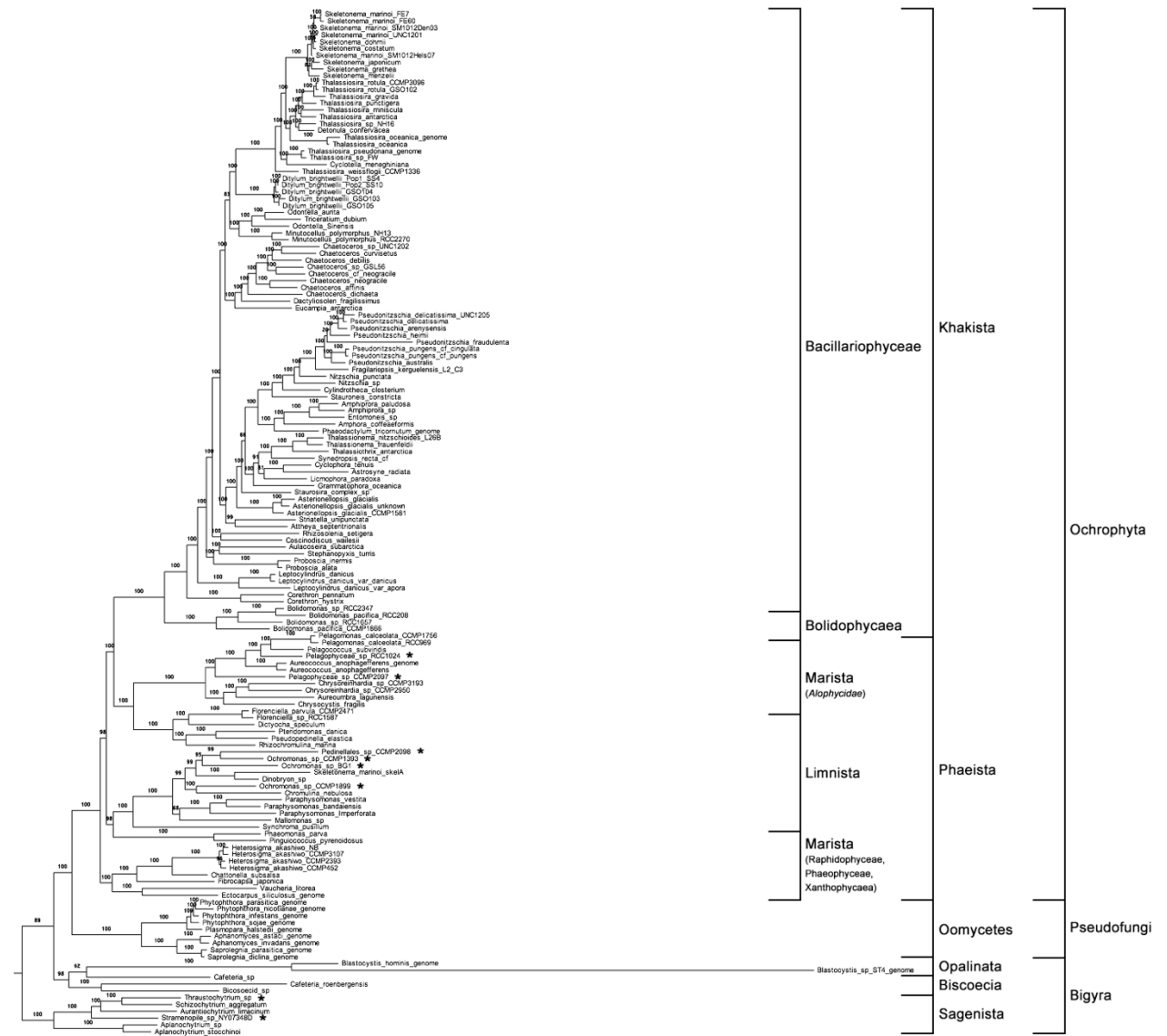
While the genes in this study were selected for their phylogenetic utility, they may also be used as a training set to provide the statistical depth to guide the selection of orthologs for genes of interest for entirely different reasons. Developing sets of genes for this use for other groups of organisms could greatly improve casual ortholog identification underlying many BLAST searches, as well as the automated annotation of genomic data.

The statistics used in the pipeline were developed to remove anomalous sequences, but they could easily be modified to select for the anomalous sequences. Anomalous sequences may be of interest in studies of horizontal gene transfer or the isolation of sequences from a symbiont or parasite.



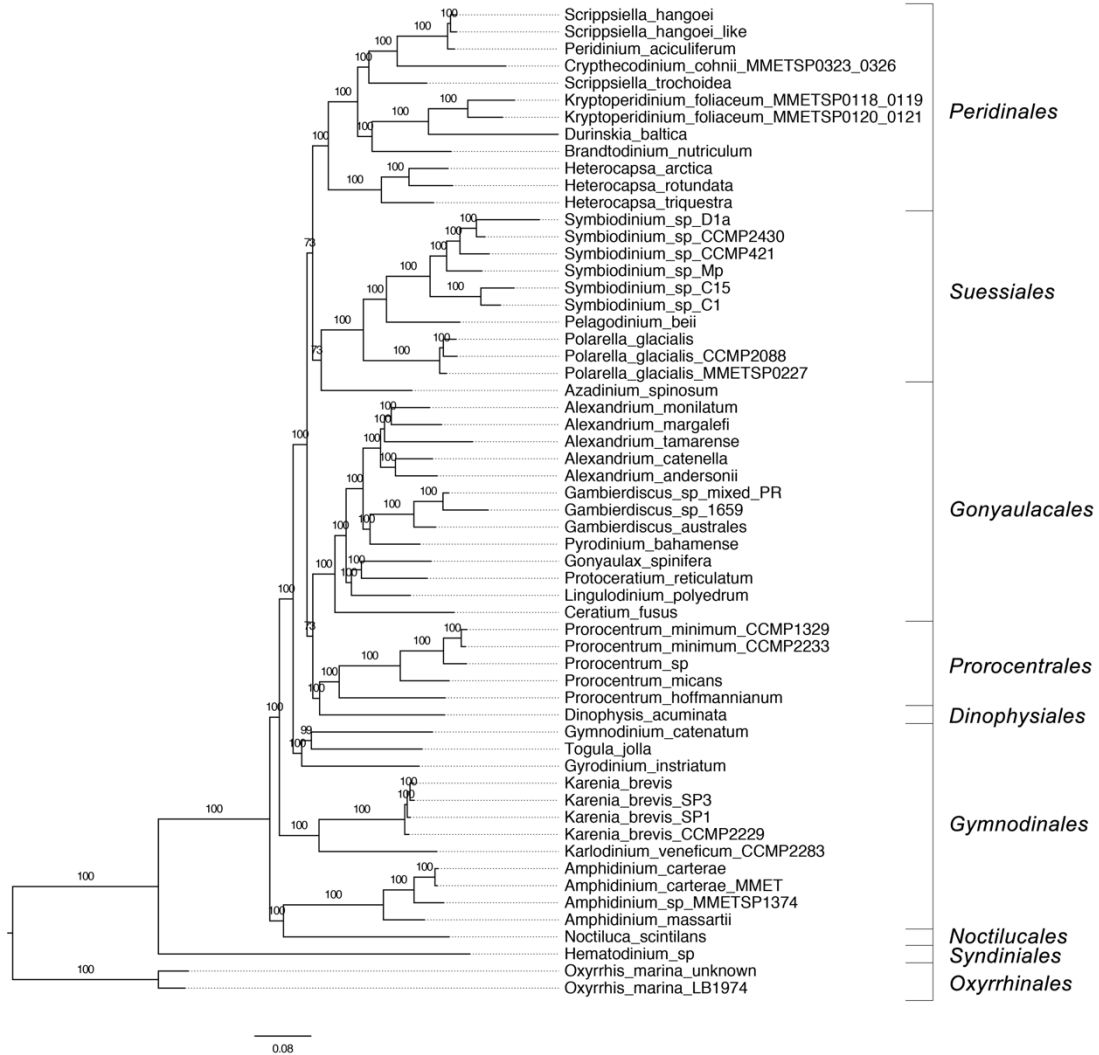
## Supplemental Figure 2 – Preliminary Stramenopile Amino Acid Maximum Likelihood Tree

The most likely tree found using RAXML using automatic selection of a substitution matrix with gamma correction and 100 bootstraps. This tree was found prior to any screening.



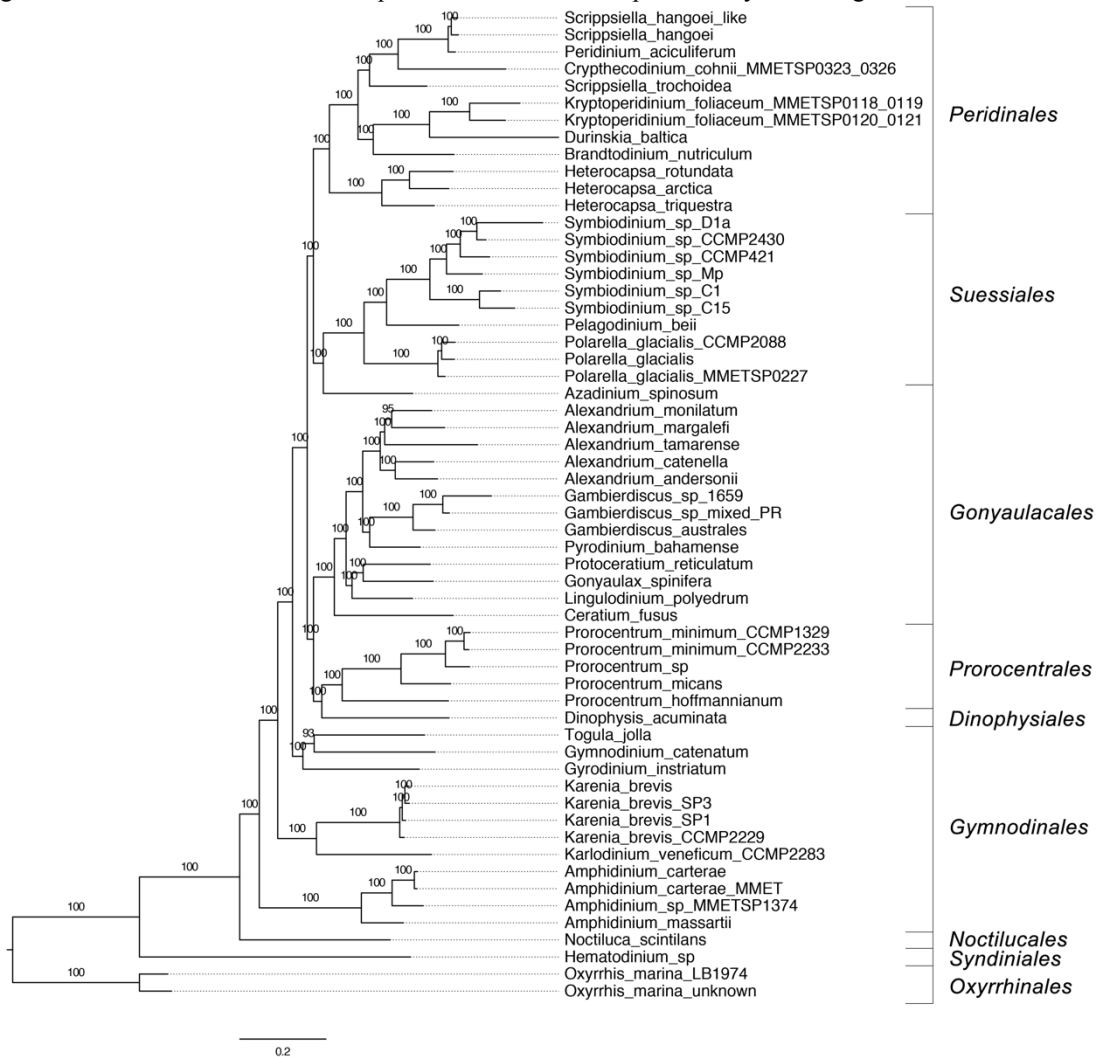
### Supplemental Figure 3 – Preliminary Dinoflagellate Nucleic Acid Maximum Likelihood Tree

The most likely tree found using RAxML using GTR substitution matrix with gamma correction and 100 bootstraps. This tree was found prior to any screening.



# Supplemental Figure 4 – Preliminary Dinoflagellate Amino Acid Maximum Likelihood Tree

The most likely tree found using RAxML using automatic selection of a substitution matrix with gamma correction and 100 bootstraps. This tree was found prior to any screening.



## Appendix B – Supplemental Tables

**Supplemental Table 1 – Organisms used in Stramenopile Case Study**

Data Type	Library ID	Strain	Name	Phylum	Class	Order
transcriptome	MMETSP0115	ms1	Bicosoecid_sp	Bicosoecida	Bicosoecida	Bicosoecida
transcriptome	MMETSP0942	E4-10	Cafeteria_roenbergensis	Bigyra	Bicosoecida	Anoecida
transcriptome	MMETSP1104	Caron Lab Isolate	Cafeteria_sp	Bigyra	Bicosoecida	Anoecida
genome	GCF_000151665.1	isolate B (sub-type 7)	Blastocystis_hominis	Bigyra	Blastocystea	Blastocystida
genome	GCF_000743755.1	WR1	Blastocystis_sp_ST4	Bigyra	Blastocystea	Blastocystida
genome	JGI 20121220		Aplanochytrium_kerguelense	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP0954-7	PBS07	Aplanochytrium_sp	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP1346-9	GSBS06	Aplanochytrium_stocchinoi	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP0958-61	ATCCMYA-1381	Aurantiochytrium_limacinum	Bigyra	Labyrinthulea	Thraustochytriida
genome	JGI 20120618		Aurantiochytrium_limacinum	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP0962-5	ATCC28209	Schizochytrium_aggregatum	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP0198,9	LLF1b	Thraustochytrium_sp	Bigyra	Labyrinthulea	Thraustochytriida
transcriptome	MMETSP1064	CCAP 1002/5	Aulacoseira_subarctica	Ochrophyta	Bacillariophyceae	Aulacoseirales
transcriptome	MMETSP0017	KMMCC:B-181	Cylindrotheca_closterium	Ochrophyta	Bacillariophyceae	Bacillariales
genome	JGI		Fragilariopsis_cylindrus	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0906-9	L2-C3	Fragilariopsis_kerguelensis_L2_C3	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0744-7	CCMP561	Nitzschia_punctata	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0014	RCC80	Nitzschia_sp	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0329	B593	Pseudo_nitzschia_arenysensis	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0139-42	10249 10 AB	Pseudo_nitzschia_australis	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0327	B596	Pseudo_nitzschia_delicatissima	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP0850-3	WWA7	Pseudo_nitzschia_fraudulenta	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP1060	cf. cingulata	Pseudo_nitzschia_pungens_cf_cingulata	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP1061	cf. pungens	Pseudo_nitzschia_pungens_cf_pungens	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP1432	UNC1205	Pseudo_nitzschia_delicatissima_UNC1205	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP1423	UNC1101	Pseudo_nitzschia_heimii	Ochrophyta	Bacillariophyceae	Bacillariales
transcriptome	MMETSP1449	CCMP2084	Attheya_septentrionalis	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0088-92	CCMP159	Chaetoceros_affinis	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP1435	CCMP164	Chaetoceros_brevis	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP1336	RCC1993	Chaetoceros_cf_neogracile	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0716-9	unknown	Chaetoceros_curvisetus	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0149-50	MM31A-1	Chaetoceros_debilis	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP1447	CCMP1751	Chaetoceros_dichaeta	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0751-4	CCMP1317	Chaetoceros_neogracile	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0200	GSL56	Chaetoceros_sp_GSL56	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP1429	UNC1202	Chaetoceros_sp_UNC1202	Ochrophyta	Bacillariophyceae	Chaetocerotanae
transcriptome	MMETSP0010	308	Corethron_hystris	Ochrophyta	Bacillariophyceae	Corethrales
transcriptome	MMETSP0169,71	L29A3	Corethron_pennatum	Ochrophyta	Bacillariophyceae	Corethrales
transcriptome	MMETSP1066	CCMP2513	Coscinodiscus_wallesii	Ochrophyta	Bacillariophyceae	Coscinodiscales
transcriptome	MMETSP0397	ECT3854	Cyclophora_tenuis	Ochrophyta	Bacillariophyceae	Cyclophorales
transcriptome	MMETSP1434	CCMP3303	Minutocellus_polymorphus_CCMP3303	Ochrophyta	Bacillariophyceae	Cymatosirales
transcriptome	MMETSP1070	NH13	Minutocellus_polymorphus_NH13	Ochrophyta	Bacillariophyceae	Cymatosirales
transcriptome	MMETSP1322	RCC2270	Minutocellus_polymorphus_RCC2270	Ochrophyta	Bacillariophyceae	Cymatosirales
transcriptome	MMETSP0705-8	CCMP134	Asterionellopsis_glacialis	Ochrophyta	Bacillariophyceae	Fragilariales
transcriptome	MMETSP1394	CCMP1581	Asterionellopsis_glacialis_CCMP1581	Ochrophyta	Bacillariophyceae	Fragilariales



transcriptome	MMETSP0713	unknown	Asterionellopsis_glacialis_unknown	Ochrophyta	Bacillariophyceae	Fragilariales
transcriptome	MMETSP1176	CCMP1620	Synedropsis_recta_cf	Ochrophyta	Bacillariophyceae	Fragilariales
transcriptome	MMETSP1437	CCMP1452	Eucampia_antarctica	Ochrophyta	Bacillariophyceae	Hemiaulales
transcriptome	MMETSP1362	CCMP1856	Leptocylindrus_danicus	Ochrophyta	Bacillariophyceae	Leptocylindrales
transcriptome	MMETSP0322	B651	Leptocylindrus_danicus_var_apora	Ochrophyta	Bacillariophyceae	Leptocylindrales
transcriptome	MMETSP0321	B650	Leptocylindrus_danicus_var_danicus	Ochrophyta	Bacillariophyceae	Leptocylindrales
transcriptome	MMETSP1360	CCMP2313	Licmophora_paradoxa	Ochrophyta	Bacillariophyceae	Licmophorales
transcriptome	MMETSP1002,5	GSO103	Ditylum_brightwellii_GSO103	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP1010,3	GSO104	Ditylum_brightwellii_GSO104	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP0998,1001	GSO105	Ditylum_brightwellii_GSO105	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP1062	Pop1 (SS4)	Ditylum_brightwellii_Pop1_SS4	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP1063	Pop2 (SS10)	Ditylum_brightwellii_Pop2_SS10	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP1171	CCMP826	Helicotheca_tamensis	Ochrophyta	Bacillariophyceae	Lithodesmiales
transcriptome	MMETSP0794	CCMP815	Stephanopyxis_turris	Ochrophyta	Bacillariophyceae	Melosirales
transcriptome	MMETSP1065	CCMP125	Amphiprora_paludosa	Ochrophyta	Bacillariophyceae	Naviculales
transcriptome	MMETSP0724-7	CCMP467	Amphiprora_sp	Ochrophyta	Bacillariophyceae	Naviculales
transcriptome	MMETSP1442	CCMP3328	Craspedostauros_australis	Ochrophyta	Bacillariophyceae	Naviculales
genome	GCF_000150955.2	CCAP 1055/1	Phaeodactylum_tricornutum	Ochrophyta	Bacillariophyceae	Naviculales
transcriptome	MMETSP1352	CCMP1120	Stauroneis_constricta	Ochrophyta	Bacillariophyceae	Naviculales
transcriptome	MMETSP1361	CCMP2646	Staurosira_complex_sp	Ochrophyta	Bacillariophyceae	Naviculales
transcriptome	MMETSP0580	unknown	Dactyliosolen_fragilissimus	Ochrophyta	Bacillariophyceae	Rhizosoleniales
transcriptome	MMETSP0174,6	PI-D3	Proboscia_alata	Ochrophyta	Bacillariophyceae	Rhizosoleniales
transcriptome	MMETSP0816	CCAP1064/1	Proboscia_inermis	Ochrophyta	Bacillariophyceae	Rhizosoleniales
transcriptome	MMETSP0789	CCMP1694	Rhizosolenia_setigera	Ochrophyta	Bacillariophyceae	Rhizosoleniales
transcriptome	MMETSP0009	CCMP 410	Grammatophora_oceanica	Ochrophyta	Bacillariophyceae	Striatellales
transcriptome	MMETSP0800	CCMP2910	Striatella_unipunctata	Ochrophyta	Bacillariophyceae	Striatellales
transcriptome	MMETSP1443	UTEXLB2267	Entomoneis_sp	Ochrophyta	Bacillariophyceae	Surirellales
transcriptome	MMETSP0786	CCMP1798	Thalassionema_frauenfeldii	Ochrophyta	Bacillariophyceae	Thalassionematales
transcriptome	MMETSP0693	unknown	Thalassionema_nitzschoides	Ochrophyta	Bacillariophyceae	Thalassionematales
transcriptome	MMETSP0156,8	L26-B	Thalassionema_nitzschoides_L26B	Ochrophyta	Bacillariophyceae	Thalassionematales
transcriptome	MMETSP0152,4	L6-D1	Thalassiothrix_antarctica	Ochrophyta	Bacillariophyceae	Thalassionematales
transcriptome	MMETSP0316-8	CCMP127	Amphora_coffeaeformis	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP1057	CCMP 338	Cyclotella_meneghiniana	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0013	1716	Skeletonema_costatum	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0562,3	SkelB	Skeletonema_dohrnii	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0578	CCMP1804	Skeletonema_grethae	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0593	CCMP2506	Skeletonema_japonicum	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP1040	FE60	Skeletonema_marinoi_FE60	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP1039	FE7	Skeletonema_marinoi_FE7	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0918,20	skelA	Skeletonema_marinoi_skelA	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0320	SM1012Den-03	Skeletonema_marinoi_SM1012Den03	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0319	SM1012Hels-07	Skeletonema_marinoi_SM1012Hels07	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP1428	UNC1201	Skeletonema_marinoi_UNC1201	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP0603,4	CCMP793	Skeletonema_menzelii	Ochrophyta	Bacillariophyceae	Thalassiosiphysales
transcriptome	MMETSP1058	CCMP353	Detonula_confervacea	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0902-5	CCMP982	Thalassiosira_antarctica	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0492-4	GMp14c1	Thalassiosira_gravida	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0737-40	CCMP1093	Thalassiosira_miniscula	Ochrophyta	Bacillariophyceae	Thalassiosirales
genome	GCA_000296195.2	CCMP1005	Thalassiosira_oceanica	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0970-3	CCMP1005	Thalassiosira_oceanica	Ochrophyta	Bacillariophyceae	Thalassiosirales
genome	GCF_000149405.2	CCMP1335	Thalassiosira_pseudonana	Ochrophyta	Bacillariophyceae	Thalassiosirales

transcriptome	MMETSP1067	Tpunct2005C2	Thalassiosira_punctigera	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0403,4	CCMP3096	Thalassiosira_rotula_CCMP3096	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0910-3	GSO102	Thalassiosira_rotula_GSO102	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP1059	FW	Thalassiosira_sp_FW	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP1071	NH16	Thalassiosira_sp_NH16	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0878-81	CCMP1336	Thalassiosira_weissflogii_CCMP1336	Ochrophyta	Bacillariophyceae	Thalassiosirales
transcriptome	MMETSP0015	isolate 1302-5	Odontella_aurita	Ochrophyta	Bacillariophyceae	Triceratiales
transcriptome	MMETSP0160	Grunow 1884	Odontella_Sinensis	Ochrophyta	Bacillariophyceae	Triceratiales
transcriptome	MMETSP1175	CCMP147	Triceratium_dubium	Ochrophyta	Bacillariophyceae	Triceratiales
transcriptome	MMETSP0418	13vi08-1A	Astrosyne_radiata	Ochrophyta	Bacillariophyceae	
transcriptome	MMETSP0785	CCMP1866	Bolidomonas_pacifica_CCMP1866	Ochrophyta	Bolidophyceae	Bolidomonadales
transcriptome	MMETSP1319	RCC208	Bolidomonas_pacifica_RCC208	Ochrophyta	Bolidophyceae	Bolidomonadales
transcriptome	MMETSP1321	RCC1657	Bolidomonas_sp_RCC1657	Ochrophyta	Bolidophyceae	Bolidomonadales
transcriptome	MMETSP1320	RCC2347	Bolidomonas_sp_RCC2347	Ochrophyta	Bolidophyceae	Bolidomonadales
transcriptome	MMETSP0019,20,0812	UTEXLB2267	Dinobryon_sp	Ochrophyta	Chrysophyceae	Chromulinales
transcriptome	MMETSP1095	UTEXLB2642	Chromulina_nebulosa	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP1105	BG1	Ochromonas_sp_BG1	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP0004-5	CCMP1393	Ochromonas_sp_CCMP1393	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP1177	CCMP1899	Ochromonas_sp_CCMP1899	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP1103	Caron Lab Isolate	Paraphysomonas_bandaiensis	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP0103-4	PA2	Paraphysomonas_imperforata	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP1107	GFlagA	Paraphysomonas_vestita	Ochrophyta	Chrysophyceae	Ochromonadales
transcriptome	MMETSP1167	CCMP3275	Mallomonas_Sp	Ochrophyta	Chrysophyceae	Synurales
transcriptome	MMETSP1174	unknown	Dictyocha_speculum	Ochrophyta	Dictyochophyceae	Dictyochales
transcriptome	MMETSP1344	CCMP2471	Florenciella_parvula_CCMP2471	Ochrophyta	Dictyochophyceae	Florenciellales
transcriptome	MMETSP1323	RCC1693	Florenciella_parvula_RCC1693	Ochrophyta	Dictyochophyceae	Florenciellales
transcriptome	MMETSP1325	RCC1007	Florenciella_sp_RCC1007	Ochrophyta	Dictyochophyceae	Florenciellales
transcriptome	MMETSP1324	RCC1587	Florenciella_sp_RCC1587	Ochrophyta	Dictyochophyceae	Florenciellales
transcriptome	MMETSP0990-3	CCMP2098	Pedinellales_sp_CCMP2098	Ochrophyta	Dictyochophyceae	Pedinellales
transcriptome	MMETSP1068,97	CCMP716	Pseudopedinella_elastica	Ochrophyta	Dictyochophyceae	Pedinellales
transcriptome	MMETSP0101,2	PT	Pteridomonas_danica	Ochrophyta	Dictyochophyceae	Pedinellales
transcriptome	MMETSP0914-7	CCMP1850	Aureococcus_anophagefferens	Ochrophyta	Dictyochophyceae	Pelagomonadales
genome	GCF_000186865.1	CCMP1984	Aureococcus_anophagefferens	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP0882-5	CCMP1429	Pelagococcus_subviridis	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP0886-9	CCMP1756	Pelagomonas_calceolata_CCMP1756	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP1328	RCC969	Pelagomonas_calceolata_RCC969	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP0974-7	CCMP2097	Pelagophyceae_sp_CCMP2097	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP1329	RCC1024	Pelagophyceae_sp_RCC1024	Ochrophyta	Dictyochophyceae	Pelagomonadales
transcriptome	MMETSP1163	CCMP2877	Phaeomonas_parva	Ochrophyta	Dictyochophyceae	Pinguiochrysidales
transcriptome	MMETSP1160	CCMP2078	Pinguicoccus_pyrenoidosus	Ochrophyta	Dictyochophyceae	Pinguiochrysidales
transcriptome	MMETSP1173	CCMP1243	Rhizochromulina_marina	Ochrophyta	Dictyochophyceae	Rhizosoleniales
transcriptome	MMETSP0890	CCMP1510	Aureoumbra_lagunensis	Ochrophyta	Dictyochophyceae	Sarcinochrysidales
transcriptome	MMETSP1165	CCMP3189	Chrysocystis_fragilis	Ochrophyta	Dictyochophyceae	Sarcinochrysidales
transcriptome	MMETSP1164	CCMP2950	Chrysoreinhardia_sp_CCMP2950	Ochrophyta	Dictyochophyceae	Sarcinochrysidales
transcriptome	MMETSP1166	CCMP3193	Chrysoreinhardia_sp_CCMP3193	Ochrophyta	Dictyochophyceae	Sarcinochrysidales
transcriptome	MMETSP1170	CCMP770	Sarcinochrysis_sp	Ochrophyta	Dictyochophyceae	Sarcinochrysidales
genome	GCF_000240725.1	CCMP526	Nannochloropsis_gaditana	Ochrophyta	Eustigmatophyceae	Eustigmatales
genome	GCA_000310025.1	Ec 32	Ectocarpus_siliculosus	Ochrophyta	Phaeophyceae	Ectocarpales
transcriptome	MMETSP0947-50	CCMP2191	Chattonella_subsalsa	Ochrophyta	Raphidophyceae	Chattonellales
transcriptome	MMETSP1339	unknown	Fibrocapsa_japonica	Ochrophyta	Raphidophyceae	Chattonellales
transcriptome	MMETSP0292-6	CCMP2393	Heterosigma_akashiwo_CCMP2393	Ochrophyta	Raphidophyceae	Chattonellales

transcriptome	MMETSP0409-11	CCMP3107	Heterosigma_akashiwo_CCMP3107	Ochrophyta	Raphidophyceae	Chattonellales
transcriptome	MMETSP0894-7	CCMP452	Heterosigma_akashiwo_CCMP452	Ochrophyta	Raphidophyceae	Chattonellales
transcriptome	MMETSP0414-6	NB	Heterosigma_akashiwo_NB	Ochrophyta	Raphidophyceae	Chattonellales
transcriptome	MMETSP1452	CCMP3072	Synchroma_pusillum	Ochrophyta	Synchromophyceae	Synchromales
transcriptome	MMETSP0945,6	CCMP2940	Vaucheria_litorea	Ochrophyta	Xanthophyceae	Vaucheriales
genome	GCA_000325885.1	LT1534	Phytophthora_capsici	Oomycota	Peronosporae	Peronosporales
genome	GCF_000142945.1	T30-4	Phytophthora_infestans	Oomycota	Peronosporae	Peronosporales
genome	GCA_001482985.1	race 1	Phytophthora_nicotianae	Oomycota	Peronosporae	Peronosporales
genome	GCF_000247585.1	INRA-310	Phytophthora_parasitica	Oomycota	Peronosporae	Peronosporales
genome	GCF_000149755.1	P6497	Phytophthora_sojae	Oomycota	Peronosporae	Peronosporales
genome	GCA_900000015.1	unknown	Plasmodium_hastidii	Oomycota	Peronosporae	Peronosporales
genome	GCF_000520075.1	AP03	Aphanomyces_astaci	Oomycota	Peronosporae	Saprolegniales
genome	GCF_000520115.1	NJM9701	Aphanomyces_invadans	Oomycota	Peronosporae	Saprolegniales
genome	GCF_000281045.1	VS20	Saprolegnia_diclina	Oomycota	Peronosporae	Saprolegniales
genome	GCF_000151545.1	CBS 223.65	Saprolegnia_parasitica	Oomycota	Peronosporae	Saprolegniales
genome	GCA_001029375.1	Pi-S	Pythium_insidiosum	Oomycota		Pythiales
transcriptome	MMETSP1433	NY07348D	Stramenopile_sp_NY07348D			

**Supplemental Table 2 – Organisms used in Dinoflagellate Study**

ID	Group	Culture Collection ID	MMET ID
Akashiwo_sanguinea	Gymnodinales	CCCM 885	MMETSP0223_2
Alexandrium_andersonii	Gonyaulacales	CCMP2222	MMETSP1436
Alexandrium_catenella	Gonyaulacales	OF101	MMETSP0790
Alexandrium_margalefi	Gonyaulacales	AMGDE01CS-322	MMETSP0661
Alexandrium_minutum	Gonyaulacales	CCMP113	MMETSP0328
Alexandrium_fundyense	Gonyaulacales	CCMP1719	MMETSP0196C MMETSP0347
Alexandrium_monilatum	Gonyaulacales	CCMP3105	MMETSP0093 MMETSP0095 MMETSP0096 MMETSP0097
Alexandrium_tamarense	Gonyaulacales	CCMP1771	MMETSP0378 MMETSP0380 MMETSP0382 MMETSP0384
Amoebophrya sp.	Syndiniales	Ameob2	MMETSP0795
Amphidinium_massartii	Gymnodinales	CS-259	MMETSP0689_2
Amphidinium_carerae	Gymnodinales	CCMP1314	MMETSP0258 MMETSP0259MMETSP0398C
Amphidinium_carerae	Gymnodinales	CCMP1314	
Amphidinium_sp	Gymnodinales		MMETSP1374
Azadinium_spinosum	Gymnodinales	3D9	MMETSP1036_2 MMETSP1037_2MMETSP1038_2
Brandtonidium_nutriculum	Peridinales	RCC3387	MMETSP1462
Ceratium_fusus	Gonyaulacales	PA161109	MMETSP1074 MMETSP1075
Chromera_velia	Chromista	CCMP2878.2	MMETSP0290
Cryptocodinium_cohnii	Gonyaulacales	Seligo	MMETSP0323_2 MMETSP0324_2 MMETSP0325_2 MMETSP0326_2
Dinophysis_acuminata	Dinophysales	DAEP01	MMETSP0797
Durinskia_baltica	Dinotrichales	CSIRO CS-38	MMETSP0116 MMETSP0117
Gambierdiscus_austales	Gonyaulacales	CAWD 149	MMETSP0766_2

Gambierdiscus_sp	Gonyaulacales	mixed_PR	
Gambierdiscus_sp	Gonyaulacales	1659	
Gonyaulax_spiniifera	Gonyaulacales	CCMP409	MMETSP1439
Gymnodinium catenatum	Gymnodinales	GC744	MMETSP0784
Gyrodinium dominans	Gymnodinales	SPMC 103	MMETSP1148
Gyrodinium_instriatum	Gymnodinales	CCMP3173	
Hematodinium_sp	Syndiniales		
Heterocapsa_arctica	Peridinales	CCMP445	MMETSP1441
Heterocapsa_rotundata	Peridinales	SCCAP K-0483	MMETSP0503
Heterocapsa_triquestra	Peridinales	CCMP 448	MMETSP0448
Karenia_brevis	Gymnodinales	Wilson	MMETSP0201 MMETSP0202
Karenia_brevis	Gymnodinales	SP3	MMETSP0527_2 MMETSP0528_2
Karenia_brevis	Gymnodinales	CCMP2229	MMETSP0027 MMETSP0029 MMETSP0030 MMETSP0031
Karenia_brevis	Gymnodinales	SP1	MMETSP0573 MMETSP0574
Karlodinium_veneficum_CCMP2283	Gymnodinales	CCMP2283	MMETSP1015 MMETSP1016 MMETSP1017
Kryptoperidinium_foliaceum_MMETSP0118_0119	Thoracosphaerales	CCAP 1116/3	MMETSP0118 MMETSP0119
Kryptoperidinium_foliaceum_MMETSP0120_0121	Dinotrichales	CCMP 1326	MMETSP0120 MMETSP0121
Lessardia_elongata	Peridinales	SPMC 104	MMETSP1147
Lingulodinium_polyedrum	Gonyaulacales	CCMP 1738	MMETSP1032 MMETSP1033 MMETSP1034 MMETSP1035
Noctiluca_scintillans	Noctilucales	unknown	MMETSP0253
Oxyrrhis_marina	Oxyrrhinales	CCMP1788	MMETSP0044
Oxyrrhis_marina	Oxyrrhinales	CCMP1795	MMETSP0451_2C
Oxyrrhis_marina	Oxyrrhinales	LB1974	MMETSP1424 MMETSP1425 MMETSP1426
Oxyrrhis_marina	Oxyrrhinales	unknown	MMETSP0468 469 470 471
Pelagodinium_bei	Suessiales	RCC1491	MMETSP1338
Peridinium_aciculiferum	Peridinales	PAER-2	MMETSP0370 MMETSP0371
Perkinsus_chesapeakei	Perkinsorida	ATCC PRA-65	MMETSP0924
Perkinsus_marinus	Perkinsorida	ATCC 50439	MMETSP0922
Perkinsus_marinus	Perkinsorida	ATCC 50983	GCA_000006405.1
Polarella_glacialis	Suessiales	CCMP1383	
Polarella_glacialis (Moore)	Suessiales	CCMP2088	MMETSP1440
Polarella_glacialis (Moore)	Suessiales	CCMP 1383	MMETSP0227
Prorocentrum_lima	Prorocentrales	CCMP 684	MMETSP0252
Prorocentrum_minimum	Prorocentrales	CCMP2233	MMETSP0267 MMETSP0268 MMETSP0269
Prorocentrum_minimum	Prorocentrales	CCMP1329	MMETSP0053 MMETSP0055 MMETSP0056 MMETSP0057
Prorocentrum_hoffmannianum	Prorocentrales	CCMP683	
Prorocentrum_micans	Prorocentrales	CCMP1589	
Prorocentrum_micans_CCCM845	Prorocentrales	CCCM 845	MMETSP0251_2
Prorocentrum_sp	Prorocentrales	CCMP3122	

Protoceratium reticulatum	Gonyaulacales	CCCM 535	MMETSP0228
Pyrocystis lunula	Pyrocystales	CCCM 517	MMETSP0229_2
Pyrodinium bahamense	Gonyaulacales	pbaha01	MMETSP0796
Scrippsiella_Hangoei	Peridinales	SHTV-5	MMETSP0359 MMETSP0360 MMETSP0361
Scrippsiella_Hangoei-like	Peridinales	SHHI-4	MMETSP0367 MMETSP0368 MMETSP0369
Scrippsiella_trochoidea	Peridinales	CCMP3099	MMETSP0270 MMETSP0271 MMETSP0272
Symbiodinium_kawagutii	Suessiales	CCMP2468	MMETSP0132_2C
Symbiodinium_sp	Suessiales	D1a	MMETSP1377
Symbiodinium_sp	Suessiales	CCMP2430	MMETSP1115 MMETSP1116 MMETSP1117
Symbiodinium_sp	Suessiales	CCMP421	MMETSP1110
Symbiodinium_sp	Suessiales	C15	MMETSP1370 MMETSP1371
Symbiodinium_sp	Suessiales	Mp	MMETSP1122 MMETSP1123 MMETSP1124 MMETSP1125
Symbiodinium_sp	Suessiales	C1	MMETSP1367 MMETSP1369
Thoracosphaera_heimii	Thoracosphaerales	CCCM 670)	MMETSP0225
Togula_jolla	Gymnodinales	CCCM 725	MMETSP0224
Vitrella_brassicaformis	Chromista	CCMP3346	MMETSP1451
Vitrella_brassicaformis_CCMP3155	Chromista	CCMP3155	MMETSP0288

## References

1. Dodge JD. Chromosome structure in the dinoflagellate and the problem of the mesokaryotic cell. The International Congress Series. London: Inter. Cong. Ser. 91\_; 1965;91: 264–265.
2. Herzog M, Maroteaux L. Dinoflagellate 17S rRNA Sequence Inferred from the Gene Sequence: Evolutionary Implications. P Natl Acad Sci Usa. 1986;83: 8644–8648.
3. Lenaers G, Scholin C, Bhaud Y, Saint-Hilaire D, Herzog M. A molecular phylogeny of dinoflagellate protists (Pyrrhophyta) inferred from the sequence of 24S rRNA divergent domains D1 and D8. J Mol Evol. 1991;32: 53–63.
4. Cavalier-Smith T. Kingdom protozoa and its 18 phyla. Microbiol Rev. 1993;57: 953–994.
5. Ki J. Nuclear 28S rDNA phylogeny supports the basal placement of *Noctiluca scintillans* (Dinophyceae; Noctilucales) in dinoflagellates. Eur J Protistol. 2010.
6. Gómez F, Moreira D, López-García P. Molecular Phylogeny of Noctiluroid Dinoflagellates (Noctilucales, Dinophyceae). Protist. 2010.
7. Gómez F, Moreira D, López-García P. *Neoceratium* gen. nov., a New Genus for All Marine Species Currently Assigned to *Ceratium* (Dinophyceae). Protist. 2010.
8. Bachvaroff T, Handy S, Delwiche C. Molecular phylogeny of ocelloid-bearing dinoflagellates (Warnowiaceae) as inferred from SSU and LSU rDNA sequences. BMC Evol Biol. 2009.
9. Handy SM, Bachvaroff TR, Timme RE, Coats DW, Kim S, Delwiche CF. Phylogeny of Four Dinophysiacean Genera (Dinophyceae, Dinophysiales) Based on rDNA Sequences From Single Cells and Environmental SAMPLES. 2009;45: 1163–1174. doi:10.1111/j.1529-8817.2009.00738.x
10. Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. Dinoflagellate expressed indicate massive transfer to the nuclear genome sequence tag data of chloroplast genes. Protist. 2004;155: 65–78.
11. Zhang Z, Cavalier-Smith T, Green BR. A family of selfish minicircular chromosomes with jumbled chloroplast gene fragments from a dinoflagellate. Mol Biol Evol. 2001;18: 1558–1565.
12. Leung SK, Wong JTY. The replication of plastid minicircles involves rolling circle intermediates. Nucleic Acids Res. 2009;37: 1991–2002.

doi:10.1093/nar/gkp063

13. LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW. Symbiodinium (Pyrrophyta) genome sizes (DNA content) are smallest among dinoflagellates. *Journal of Phycology*. Blackwell Science Inc; 2005;41: 880–886. doi:10.1111/j.0022-3646.2005.04231.x
14. Roberts T, Tuttle R, Allen J, Loeblich A, KLOTZ L. New Genetic and Physicochemical Data on Structure of Dinoflagellate Chromosomes. *Nature*. 1974;248: 446–447.
15. Yoshikawa T, Uchida A, Ishida Y. There are 4 introns in the gene coding the DNA-binding protein HCC of *Cryptocodinium cohnii* (Dinophyceae). *Fisheries Sci.* 1996;62: 204–209.
16. Le QH, Markovic P, Hastings JW, Jovine RV, Morse D. Structure and organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*. *Mol Gen Genet.* 1997;255: 595–604.
17. Rowan R, Whitney S, Fowler A, Yellowlees D. Rubisco in marine symbiotic dinoflagellates: Form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. *Plant Cell.* 1996;8: 539–553.
18. Machabee S, Wall L, Morse D. Expression and genomic organization of a dinoflagellate gene family. *Plant Mol Biol.* 1994;25: 23–31.
19. Bachvaroff TR, Place AR. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. *PLoS ONE.* 2008;3: e2929.
20. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The Symbiodinium *kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *American Association for the Advancement of Science*; 2015;350: 691–694. doi:10.1126/science.aad0408
21. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the Symbiodinium *minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol. Elsevier*; 2013;23: 1399–1408. doi:10.1016/j.cub.2013.05.062
22. Beedessee G, Hisata K, Roy MC, Satoh N, Shoguchi E. Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, *Symbiodinium minutum*. *Bmc Genomics. BioMed Central*; 2015;16: 941. doi:10.1186/s12864-015-2195-8
23. Taylor FJR. *The Biology of Dinoflagellates*. Blackwell Scientific Publications; 1987.

24. Taylor FJR, Hoppenrath M, Saldarriaga JF. Dinoflagellate diversity and distribution. *Biodivers Conserv.* 2008;17: 407–418. doi:10.1007/s10531-007-9258-3
25. Fensome R, Saldarriaga J, Taylor F. Dinoflagellate phylogeny revisited: reconciling morphological and molecular based phylogenies. *Grana.* 1999;38: 66–80.
26. Miller JJ, Delwiche CF, Coats DW. Ultrastructure of *Amoebophrya* sp. and its changes during the course of infection. *Protist.* 2012;163: 720–745. doi:10.1016/j.protis.2011.11.007
27. Fensome RA, Saldarriaga JF, Taylor MFJR. Dinoflagellate phylogeny revisited: reconciling morphological and molecular based phylogenies. *Grana. The biology of dinoflagellates*; 1999;38: 66–80. doi:10.1080/00173139908559216
28. Bachvaroff TR, Gornik SG, Concepcion GT, Waller RF, Mendez GS, Lippmeier JC, et al. Dinoflagellate phylogeny revisited: using ribosomal proteins to resolve deep branching dinoflagellate clades. *Mol Phylogenet Evol.* 2014;70: 314–322. doi:10.1016/j.ympev.2013.10.007
29. Rae P. 5-Hydroxymethyluracil in the DNA of a dinoflagellate. *P Natl Acad Sci Usa.* 1973;70: 1141.
30. Rill RL, Livolant F, Aldrich HC, Davidson MW. Electron microscopy of liquid crystalline DNA: direct evidence for cholesteric-like organization of DNA in dinoflagellate chromosomes. *Chromosoma.* 1989;98: 280–286.
31. Kim S, Bachvaroff TR, Handy SM, Delwiche CF. Dynamics of actin evolution in dinoflagellates. *Mol Biol Evol.* 2011;28: 1469–1480. doi:10.1093/molbev/msq332
32. Zhang H, Hou Y, Lin S. Isolation and characterization of proliferating cell nuclear antigen from the dinoflagellate *Pfiesteria piscicida*. *J Eukaryot Microbiol.* 2006;53: 142–150. doi:10.1111/j.1550-7408.2005.00085.x
33. Gajadhar A, Marquardt W, Hall R, Gunderson J, Ariztia-Carmona E, Sogin M. Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Cryptocodium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. *Molecular and biochemical parasitology.* 1991;45: 147–154.
34. Parrow M, Elbrächter M, Krause M, Burkholder J, Deamer N, Htaye N, et al. The taxonomy and growth of a *Cryptocodium* species (Dinophyceae) isolated from a brackish-water fish aquarium. *African Journal of Marine Science.* 2006;28: 185–191.



35. Logares R, Shalchian-Tabrizi K, Boltovskoy A, Rengefors K. Extensive dinoflagellate phylogenies indicate infrequent marine-freshwater transitions. *Mol Phylogenet Evol.* 2007;45: 887–903. doi:10.1016/j.ympev.2007.08.005
36. Saldarriaga JF, Cavalier-Smith T. Molecular data and the evolutionary history of dinoflagellates. *Eur J Protistol.* 2004;40: 85–111. doi:10.1016/j.ejop.2003.11.003
37. Saldarriaga JF, Taylor FJ, Keeling PJ, Cavalier-Smith T. Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J Mol Evol.* Springer-Verlag; 2001;53: 204–213. doi:10.1007/s002390010210
38. Lin S, Zhang H, Jiao N. Potential utility of mitochondrial cytochrome B and its mRNA editing in resolving closely related dinoflagellates: a case study of *Prorocentrum* (Dinophyceae). *Journal of Phycology.* Blackwell Publishing Inc; 2006;42: 646–654. doi:10.1111/j.1529-8817.2006.00229.x
39. Zhang H, Bhattacharya D, Lin S. Phylogeny of dinoflagellates based on mitochondrial Cytochrome b and nuclear small subunit rDNA sequence comparisons. *Journal of Phycology.* Blackwell Science Inc; 2005;41: 411–420. doi:10.1111/j.1529-8817.2005.04168.x
40. Harper JT, Waanders E, Keeling PJ. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int J Syst Evol Micr.* Society for General Microbiology; 2005;55: 487–496. doi:10.1099/ijls.0.63216-0
41. Wynn J, Behrens P, Sundararajan A, Hansen J, Apt K. Production of single cell oils by dinoflagellates. *Single cell oils.* AOCS Press; 2005.
42. Allen J, Roberts T, Loeblich A. Characterization of the DNA from the dinoflagellate *Cryptothecodinium cohnii* and implications for nuclear organization. *Cell.* 1975;6: 161–169.
43. Xue L, Lippmeier JC, Bingham S, Apt KE. ESTs from the dinoflagellate *Cryptothecodinium cohnii*. Baltimore; 1999.
44. Lee DH, Mittag M, Sczekan S, Morse D, Hastings JW. Molecular cloning and genomic organization of a gene for luciferin-binding protein from the dinoflagellate *Gonyaulax polyedra*. *J Biol Chem.* 1993;268: 8842–8850.
45. Sharples FP, Wrench PM, Ou K, Hiller RG. Two distinct forms of the peridinin-chlorophyll a-protein from *Amphidinium carterae*. *Biochim Biophys Acta.* 1996;1276: 117–123.
46. Hiller RG, Crossley LG, Wrench PM, Santucci N, Hofmann E. The 15-kDa forms of the apo-peridinin-chlorophyll a protein (PCP) in dinoflagellates show high identity with the apo-32 kDa PCP forms, and have similar N-terminal

- leaders and gene arrangements. *Mol Genet Genomics*. 2001;266: 254–259.
47. Li L, Hastings JW. The structure and organization of the luciferase gene in the photosynthetic dinoflagellate *Gonyaulax polyedra*. *Plant Mol Biol*. 1998;36: 275–284.
  48. Reichman JR, Wilcox TP, Vize PD. PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low levels of concerted evolution, isoform diversity, and spectral tuning of chromophores. *Mol Biol Evol*. 2003;20: 2143–2154. doi:10.1093/molbev/msg233
  49. Okamoto OK, Liu L, Robertson DL, Hastings JW. Members of a dinoflagellate luciferase gene family differ in synonymous substitution rates. *Biochemistry*. 2001;40: 15862–15868.
  50. Zhang H, Lin S. Complex gene structure of the form ii rubisco in the dinoflagellate *Prorocentrum minimum* (dinophyceae). *Journal of Phycology*. Blackwell Science Inc; 2003;39: 1160–1171. doi:10.1111/j.0022-3646.2003.03-055.x
  51. Lidie K, van Dolah F. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. *J Eukaryot Microbiol*. 2007;54: 427–435.
  52. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, et al. Spliced leader RNA trans-splicing in dinoflagellates. *P Natl Acad Sci Usa*. 2007;104: 4618–4623. doi:10.1073/pnas.0700258104
  53. Zhang MQ. Statistical features of human exons and their flanking regions. *Hum Mol Genet*. 1998;7: 919–932.
  54. Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res*. 1992;20: 4255–4262.
  55. Wu Q, Krainer AR. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol Cell Biol*. 1999;19: 3225–3236.
  56. Thanaraj TA, Clark F. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res*. 2001;29: 2581–2593.
  57. Burge CB, Padgett RA, Sharp PA. Evolutionary fates and origins of U12-type introns. *Mol Cell*. 1998;2: 773–785.
  58. Tarn WY, Steitz JA. Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem Sci*. 1997;22: 132–137. doi:10.1016/S0968-0004(97)01018-9

59. Okamoto OK, Robertson DL, Fagan TF, Hastings JW, Colepiccolo P. Different regulatory mechanisms modulate the expression of a dinoflagellate iron-superoxide dismutase. *J Biol Chem*. 2001;276: 19989–19993. doi:10.1074/jbc.M101169200
60. Sambrook J, Russell DW. *Molecular Cloning*. 3rd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2001.
61. McEwan M, Humayun R, Slamovits CH, Keeling PJ. Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. *J Eukaryot Microbiol*. Blackwell Publishing Inc; 2008;55: 530–535. doi:10.1111/j.1550-7408.2008.00357.x
62. Slamovits CH, Keeling PJ. Widespread recycling of processed cDNAs in dinoflagellates. *Curr Biol*. Elsevier; 2008;18: R550–2. doi:10.1016/j.cub.2008.04.054
63. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. Oxford University Press; 2011;39: W29–37. doi:10.1093/nar/gkr367
64. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. Oxford University Press; 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351
65. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;23: 1061–1067. doi:10.1093/bioinformatics/btm071
66. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13: 2178–2189. doi:10.1101/gr.1224503
67. Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, Abdellateef M, et al. SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. *BMC Bioinformatics*. BioMed Central Ltd; 2014;15: 229. doi:10.1186/1471-2105-15-229
68. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*. Oxford University Press; 2004;32: W309–12. doi:10.1093/nar/gkh379
69. Dunn CW, Howison M, Zapata F. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*. BioMed Central Ltd; 2013;14: 330. doi:10.1186/1471-2105-14-330
70. Jessica R Grant LAK. Building a Phylogenomic Pipeline for the Eukaryotic

Tree of Life - Addressing Deep Phylogenies with Genome-Scale Data. PLoS Currents. Public Library of Science; 2014;6.  
doi:10.1371/currents.tol.c24b6054aebf3602748ac042ccc8f2e9

71. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. Oxford University Press; 2010;26: 2460–2461.  
doi:10.1093/bioinformatics/btq461
72. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 2009;23: 205–211.
73. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. Oxford University Press; 2013;30: 772–780. doi:10.1093/molbev/mst010
74. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. Oxford University Press; 2009;25: 1972–1973.  
doi:10.1093/bioinformatics/btp348
75. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. Oxford University Press; 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033
76. Iglewicz B, Hoaglin D. Volume 16: How to Detect and Handle Outliers, in *The ASQC Basic References in Quality Control: Statistical Techniques*. ASQC Quality Press; 1993.
77. Shapiro BJ, Alm EJ. Comparing patterns of natural selection across species using selective signatures. *PLoS Genet*. Public Library of Science; 2008;4: e23. doi:10.1371/journal.pgen.0040023
78. Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep*. Nature Publishing Group; 2014;4: 6504. doi:10.1038/srep06504
79. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. Roberts RG, editor. *PLoS Biol*. Public Library of Science; 2014;12: e1001889.  
doi:10.1371/journal.pbio.1001889
80. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7: 562–578. doi:10.1038/nprot.2012.016
81. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.

- Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*. NIH Public Access; 2011;29: 644–652. doi:10.1038/nbt.1883
82. Roure B, Rodriguez-Ezpeleta N, Philippe H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *Bmc Evol Biol*. BioMed Central; 2007;7 Suppl 1: S2. doi:10.1186/1471-2148-7-S1-S2
  83. Delwiche CF, Anderson RA, Bhattacharya D, Mishler B, Richard MM. Algal evolution and the early radiation of green plants. In: Cracraft J, Donoghue MJ, editors. *Assembling the Tree of Life*. London; 2004. pp. 121–137.
  84. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. *Science*. American Association for the Advancement of Science; 2015;348: –1261605. doi:10.1126/science.1261605
  85. Murray S, Ip CLC, Moore R, Nagahama Y, Fukuyo Y. Are Prorocentroid Dinoflagellates Monophyletic? A Study of 25 Species Based on Nuclear and Mitochondrial Genes. *Protist*. 2009;160: 245–264. doi:10.1016/j.protis.2008.12.004
  86. Lin S, Zhang H, Hou Y, Zhuang Y. High-Level Diversity of Dinoflagellates in the Natural Environment, Revealed by Assessment of Mitochondrial *cox1* and *cob* Genes for Dinoflagellate DNA Barcoding. *Applied and ....* 2009.
  87. Fukuda Y, Endoh H. Phylogenetic analyses of the dinoflagellate *Noctiluca scintillans* based on [ $\beta$ ]-tubulin and Hsp90 genes. *Eur J Protistol*. 2008.
  88. Gornik SG, Febrimarsa, Cassin AM, MacRae JI, Ramaprasad A, Rchiad Z, et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *P Natl Acad Sci Usa*. National Acad Sciences; 2015;112: 5767–5772. doi:10.1073/pnas.1423400112
  89. Contreras-Moreira B, Vinuesa P. GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*. American Society for Microbiology; 2013;79: 7696–7701. doi:10.1128/AEM.02411-13
  90. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*. Oxford University Press; 2010;26: 1481–1487. doi:10.1093/bioinformatics/btq229
  91. Zhang H, Bhattacharya D, Lin S. A three-gene dinoflagellate phylogeny suggests monophyly of prorocentrales and a basal position for amphidinium and heterocapsa. *J Mol Evol*. 2007;65: 463–474. doi:10.1007/s00239-007-9038-4

92. Fensome RA, History AMON. A Classification of Living and Fossil Dinoflagellates. American Museum of Natural History; 1993.
93. Logares R, Rengefors K, Kremp A, Shalchian-Tabrizi K, Boltovskoy A, Tengs T, et al. Phenotypically Different Microalgal Morphospecies with Identical Ribosomal DNA: A Case of Rapid Adaptive Evolution? *Microb Ecol.* Springer-Verlag; 2007;53: 549–561. doi:10.1007/s00248-006-9088-y
94. Taylor F. Illumination or confusion? Dinoflagellate molecular phylogenetic data viewed from a primarily morphological standpoint. *Phycol Res.* 2004;52: 308–324.